

# Colored stochastic dominance problems

Jie Xue\*  
xuexx193@umn.edu

Yuan Li\*  
lix2100@umn.edu

## Abstract

In this paper, we study the dominance relation under a stochastic setting. Let  $\mathcal{S}$  be a set of  $n$  colored stochastic points in  $\mathbb{R}^d$ , each of which is associated with an existence probability. We investigate the problem of computing the probability that a realization of  $\mathcal{S}$  contains inter-color dominances, which we call the *colored stochastic dominance* (CSD) problem. We propose the first algorithm to solve the CSD problem for  $d = 2$  in  $O(n^2 \log^2 n)$  time. On the other hand, we prove that, for  $d \geq 3$ , even the CSD problem with a restricted color pattern is  $\#P$ -hard. In addition, even if the existence probabilities are restricted to be  $\frac{1}{2}$ , the problem remains  $\#P$ -hard for  $d \geq 7$ . A simple FPRAS is then provided to approximate the desired probability in any dimension. We also study a variant of the CSD problem in which the dominance relation is considered with respect to not only the standard basis but any orthogonal basis of  $\mathbb{R}^d$ . Specifically, this variant, which we call the *free-basis colored stochastic dominance* (FBCSD) problem, considers the probability that a realization of  $\mathcal{S}$  contains inter-color dominances with respect to any orthogonal basis of  $\mathbb{R}^d$ . We show that the CSD problem is polynomial-time reducible to the FBCSD problem in the same dimension, which proves the  $\#P$ -hardness of the latter for  $d \geq 3$ . Conversely, we reduce the FBCSD problem in  $\mathbb{R}^2$  to the CSD problem in  $\mathbb{R}^2$ , by which an  $O(n^4 \log^2 n)$  time algorithm for the former is obtained.

## 1 Introduction

A point  $p \in \mathbb{R}^d$  is said to *dominate* another point  $q \in \mathbb{R}^d$  (denoted by  $p > q$ ) if the coordinate of  $p$  is greater than or equal to the coordinate of  $q$  in every dimension. The dominance relation is an important notion in multi-criteria decision-making, and has been well-studied in computational geometry, database, optimization, and other related areas. In the last decades, many problems regarding dominance relation have been proposed and investigated, e.g., deciding the dominance-existence of a dataset, counting the number of dominance pairs, reporting the dominance pairs, etc.

The dominance relation is traditionally studied on certain datasets, in which the information of each data point is exactly known. However, in many real-world applications, due to noise and limitation of devices, the data obtained may be imprecise or not totally reliable. In this situation, uncertain datasets (or stochastic datasets), in which the data points are allowed to have some uncertainties, can better capture the features of real data, and is thus more preferable than certain ones. Motivated by this, in this paper, we investigate several problems regarding the dominance relation on stochastic datasets. Following [4, 5, 7, 8], the stochastic datasets to be considered in this paper are provided with existential uncertainty: each data point has a certain (known) location but an uncertain existence depicted by an associated existence probability.

Given such a stochastic dataset in  $\mathbb{R}^d$ , there is a natural question regarding the dominance relation: how likely does a realization of the dataset (i.e., a random sample of the points with the corresponding existence probabilities) contain dominance pairs? This question can be formulated as the problem of computing the probability that a realization contains dominance pairs. Also, one may consider a bichromatic version of the problem: the given stochastic dataset consists of bichromatic (say red and blue) points and when computing the probability, the dominance pairs of interest are those formed by one red point and one blue

---

\*Dept. of Computer Science and Engg., Univ. of Minnesota — Twin Cities, 4-192 Keller Hall, 200 Union St. SE, Minneapolis, MN 55455, USA

point (similarly to the red/blue dominance problems in conventional setting). In this paper, we combine and generalize the above two problems into a so-called *colored stochastic dominance* (CSD) problem. Let  $\mathcal{S}$  be a colored stochastic dataset in  $\mathbb{R}^d$  (i.e., each data point is associated with a color and an existence probability). The CSD problem considers the probability that a realization of  $\mathcal{S}$  contains inter-color dominances, i.e., dominance pairs formed by two points of different colors.

The dominance relation is not preserved even under isometric transformations of  $\mathbb{R}^d$ . However, feature transformations are in fact commonly used for analyzing complex data in many real-world applications. Therefore, in some situations, people are interested in the dominance relation not only in the original feature space but also under (proper) transformations of the feature space. For example, given a set of points in  $\mathbb{R}^d$ , one may ask whether there always exists dominance pairs under any isometric transformation of  $\mathbb{R}^d$ , or equivalently, with respect to any orthogonal basis of  $\mathbb{R}^d$  (see Section 3 for a formal definition). Motivated by this, we also investigate a variant of the CSD problem, which we call the *free-basis colored stochastic dominance* (FBCSD) problem. Let  $\mathcal{S}$  be a colored stochastic dataset in  $\mathbb{R}^d$ . The FBCSD problem considers the probability that a realization of  $\mathcal{S}$  contains inter-color dominances with respect to any orthogonal basis of  $\mathbb{R}^d$ .

**Our results.** ( $n$  is the number of the points and  $d$  is the dimension which is fixed.)

1. We solve the CSD problem for  $d = 2$  in  $O(n^2 \log^2 n)$  time.
2. We prove that, for  $d \geq 3$ , even the CSD problem with a restricted color pattern is #P-hard. Furthermore, even if the existence probabilities of the points are restricted to be  $\frac{1}{2}$ , the problem remains #P-hard for  $d \geq 7$ .
3. We provide a simple FPRAS for the CSD problem in any dimension.
4. We show that the CSD problem is polynomial-time reducible to the FBCSD problem in the same dimension, which implies the #P-hardness of the latter for  $d \geq 3$ .
5. We solve the FBCSD problem for  $d = 2$  in  $O(n^4 \log^2 n)$  time.

**Related work.** The classical study regarding the dominance relation can be found in many works such as [6, 9, 10]. Recently, there have been a few works considering the dominance relation on stochastic datasets [1, 11, 19]; however, their main focus are the skyline (or dominance-maxima) problems under locational uncertainty, which is quite different from the problems studied in this paper. Besides problems regarding the dominance relation, many other fundamental geometric problems have also been investigated under stochastic settings in recent years, e.g., closest pair [8], minimum spanning tree [7], convex hull [3, 14], linear separability [5, 18], nearest neighbor search [2, 13], range-max query [4], etc.

**Basic notions and preliminaries.** We give the formal definitions of some basic notions used in this paper. A *colored stochastic dataset*  $\mathcal{S}$  in  $\mathbb{R}^d$  is represented by a 3-tuple  $\mathcal{S} = (S, \text{cl}, \pi)$ , where  $S \subset \mathbb{R}^d$  is the point set,  $\text{cl} : S \rightarrow \mathbb{N}$  is the coloring (or coloring function) indicating the colors of the points, and  $\pi : S \rightarrow [0, 1]$  is the function indicating the existence probabilities of the points, i.e., each point  $a \in S$  has the color (label)  $\text{cl}(a)$  and the existence probability  $\pi(a)$ . A *realization* of  $\mathcal{S}$  refers to a random sample  $R \subseteq S$  where each point  $a \in S$  is sampled with probability  $\pi(a)$ . For any  $A \subseteq S$ , an *inter-color dominance* in  $A$  with respect to  $\text{cl}$  (or simply an inter-color dominance in  $A$  if the coloring  $\text{cl}$  is unambiguous) refers to a pair  $(a, b)$  with  $a, b \in A$  such that  $\text{cl}(a) \neq \text{cl}(b)$  and  $a > b$ . A *sub-dataset* of  $\mathcal{S}$  is a colored stochastic dataset  $\mathcal{S}' = (S', \text{cl}', \pi')$  where  $S' \subseteq S$ ,  $\text{cl}' = \text{cl}|_{S'}$ ,  $\pi' = \pi|_{S'}$ . A *bipartite* graph is represented as  $G = (V \cup V', E)$ , where  $V, V'$  are the two parts (of vertices) and  $E$  is the edge set.

**All the detailed proofs of our lemmas and theorems are presented in Appendix A.**

## 2 The colored stochastic dominance problem

Let  $\mathcal{S} = (S, \text{cl}, \pi)$  be a colored stochastic dataset in  $\mathbb{R}^d$  with  $S = \{a_1, \dots, a_n\}$ . Define  $\Lambda_{\mathcal{S}}$  as the probability that a realization of  $\mathcal{S}$  contains inter-color dominances. Set  $\Gamma_{\mathcal{S}} = 1 - \Lambda_{\mathcal{S}}$ , which is the probability that a realization of  $\mathcal{S}$  contains no inter-color dominances. The goal of the CSD problem is to compute  $\Lambda_{\mathcal{S}}$  (or  $\Gamma_{\mathcal{S}}$ ).

## 2.1 An algorithm for $d = 2$

The naïve method for solving the CSD problem is to enumerate all subsets of  $S$  and “count” those containing inter-color dominances. However, it requires exponential time, as there are  $2^{|S|}$  subsets of  $S$  to be considered. In this section, we show that the CSD problem in  $\mathbb{R}^2$  can be solved much more efficiently. Specifically, we propose an  $O(n^2 \log^2 n)$ -time algorithm to compute  $\Gamma_S$ . For simplicity, we assume that the points in  $S$  have distinct  $x$ -coordinates and  $y$ -coordinates (if this is not the case, we can first “regularize”  $S$  by Lemma 15).

When computing  $\Gamma_S$ , we need to consider the realizations which contain no inter-color dominances. As we will see, in the case of  $d = 2$ , these realizations have good properties, which allows us to solve the problem efficiently in a recursive way. For any point  $a \in \mathbb{R}^2$ , we use  $x(a)$  (resp.,  $y(a)$ ) to denote the  $x$ -coordinate (resp.,  $y$ -coordinate) of  $a$ . Suppose the points  $a_1, \dots, a_n \in S$  are already sorted such that  $x(a_1) < \dots < x(a_n)$ . For convenience of exposition, we add a dummy point  $a_0$  to  $S$  with  $x(a_0) < x(a_1)$  and  $y(a_0) > y(a_i)$  for all  $i \in \{1, \dots, n\}$ . The color  $\text{cl}(a_0)$  is defined to be different from  $\text{cl}(a_1), \dots, \text{cl}(a_n)$ , and  $\pi(a_0) = 1$ . Note that including  $a_0$  does not change  $\Gamma_S$ . For a subset  $A = \{a_{i_1}, \dots, a_{i_r}\}$  of  $S$  with  $i_1 < \dots < i_r$ , we define  $Z(A) = \emptyset$  if  $A$  is monochromatic, and otherwise  $Z(A) = \{a_{i_1}, \dots, a_{i_i}\}$  such that  $\text{cl}(a_{i_i}) \neq \text{cl}(a_{i_{i+1}}) = \dots = \text{cl}(a_{i_r})$ . In other words,  $Z(A)$  is the subset of  $A$  obtained by dropping the “rightmost” points of the same color as  $a_{i_r}$ ; see Figure 1. We have the following important observation.

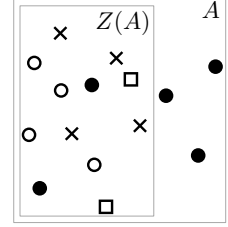


Figure 1: Illustrating  $A$  and  $Z(A)$ .

**Lemma 1** *A realization  $R$  of  $\mathcal{S}$  contains no inter-color dominances iff  $Z(R)$  contains no inter-color dominances and for any  $a \in Z(R)$ ,  $b \in R \setminus Z(R)$  it holds that  $y(a) > y(b)$ .*

With this in hand, we then consider how to compute  $\Gamma_S$ . For a nonempty subset  $A \subseteq S$ , we define the *signature*  $\text{sgn}(A)$  of  $A$  as a pair  $(i, j)$  such that  $a_i, a_j \in A$  and  $a_i$  (resp.,  $a_j$ ) has the greatest  $x$ -coordinate (resp., smallest  $y$ -coordinate) among all points in  $A$ . Let  $E_{i,j}$  be the event that a realization  $R$  of  $\mathcal{S}$  contains no inter-color dominances and satisfies  $\text{sgn}(R) = (i, j)$ . Note that if a realization  $R$  contains no inter-color dominances, then either  $R = \{a_0\}$  or some  $E_{i,j}$  happens for  $i, j \in \{1, \dots, n\}$ . So we immediately have

$$\Gamma_S = \prod_{i=1}^n (1 - \pi(a_i)) + \sum_{i=1}^n \sum_{j=1}^n \Pr[E_{i,j}].$$

Now the problem is reduced to computing all  $\Pr[E_{i,j}]$ . Instead of working on the events  $\{E_{i,j}\}$  directly, we consider a set of slightly different events  $\{E'_{i,j}\}$  defined as follows. For  $p \in \{0, \dots, n\}$ , set  $S_p = \{a_0, \dots, a_p\}$ , and we use  $\mathcal{S}_p$  to denote the sub-dataset of  $\mathcal{S}$  with point set  $S_p \subseteq S$ . Define  $E'_{i,j}$  as the event that a realization  $R$  of  $\mathcal{S}_i$  contains no inter-color dominances and satisfies  $\text{sgn}(R) = (i, j)$ . It is quite easy to see the equations

$$\Pr[E_{i,j}] = \Pr[E'_{i,j}] \cdot \prod_{t=i+1}^n (1 - \pi(a_t)).$$

Set  $F(i, j) = \Pr[E'_{i,j}]$ . We show how to compute all  $F(i, j)$  recursively by applying Lemma 1. Observe that  $F(i, j) = 0$  if  $x(a_i) < x(a_j)$  (equivalently,  $i < j$ ) or  $y(a_i) < y(a_j)$  or  $\text{cl}(a_i) \neq \text{cl}(a_j)$ . Thus, it suffices to compute all  $F(i, j)$  with  $i \geq j$ ,  $y(a_i) \geq y(a_j)$ ,  $\text{cl}(a_i) = \text{cl}(a_j)$  (we say the pair  $(i, j)$  is *legal* if these three conditions hold). Let  $(i, j)$  be a legal pair. Trivially, for  $i = j = 0$ , we have  $F(i, j) = 1$ . So suppose  $i, j > 0$ . Let  $R$  be a realization of  $\mathcal{S}_i$ . To compute  $F(i, j)$ , we consider the signature  $\text{sgn}(Z(R))$  under the condition that  $E'_{i,j}$  happens. First, when  $E'_{i,j}$  happens, we always have  $Z(R) \neq \emptyset$ , because  $R$  at least contains  $a_0, a_i, a_j$  (possibly  $i = j$ ) and  $\text{cl}(a_0) \neq \text{cl}(a_i) = \text{cl}(a_j)$ . Therefore, in this case,  $\text{sgn}(Z(R))$  is defined and must be a legal pair  $(i', j')$  for some  $i', j' \in \{0, \dots, i-1\}$ . It follows that  $F(i, j)$  can be computed by considering for each such pair  $(i', j')$  the probability that  $R$  contains no inter-color dominances and  $\text{sgn}(R) = (i, j)$ ,  $\text{sgn}(Z(R)) = (i', j')$ , and then summing up these probabilities. Note that if  $\text{sgn}(R) = (i, j)$  and  $\text{sgn}(Z(R)) = (i', j')$ , then  $i' < j$  and  $\text{cl}(i') \neq \text{cl}(i)$ . In addition, if  $R$  contains no inter-color dominances, then we must have  $y(a_i) < y(a_{j'})$  by Lemma 1. As such, we only need to consider the legal pairs  $(i', j')$

satisfying  $i' < j$ ,  $y(a_i) < y(a_{j'})$ ,  $\text{cl}(i') \neq \text{cl}(i)$  (we denote the set of these pairs by  $J_{i,j}$ ). Fixing such a pair  $(i', j') \in J_{i,j}$ , we investigate the corresponding probability. By the definition of  $Z(R)$  and Lemma 1, we observe that if  $R$  contains no inter-color dominances and  $\text{sgn}(R) = (i, j)$ ,  $\text{sgn}(Z(R)) = (i', j')$ , then

- $R \cap S_{i'}$  contains no inter-color dominances and  $\text{sgn}(R \cap S_{i'}) = (i', j')$ ;
- $R \cap (S_i \setminus S_{i'})$  includes  $a_i$  and  $a_j$ , but does not include any point  $a_t$  for  $t \in \{i' + 1, \dots, i\}$  satisfying  $\text{cl}(a_t) \neq \text{cl}(a_i)$  or  $y(a_t) < y(a_j)$  or  $y(a_{j'}) < y(a_t)$ .

Conversely, one can also verify that if a realization  $R$  of  $S_i$  satisfies the above two conditions, then  $R$  contains no inter-color dominances (by Lemma 1) and  $\text{sgn}(R) = (i, j)$ ,  $\text{sgn}(Z(R)) = (i', j')$  (note that  $Z(R) = R \cap S_{i'}$ ). Therefore, the probability that  $R$  contains no inter-color dominances and  $\text{sgn}(R) = (i, j)$ ,  $\text{sgn}(Z(R)) = (i', j')$  is just the product  $F(i', j') \cdot \pi_{i,j}^* \cdot \Pi_{i,j,i',j'}$ , where  $\pi_{i,j}^* = \pi(a_i) \cdot \pi(a_j)$  if  $i \neq j$  and  $\pi_{i,j}^* = \pi(a_i)$  if  $i = j$ , and  $\Pi_{i,j,i',j'}$  is the product of all  $(1 - \pi(a_t))$  for  $t \in \{i' + 1, \dots, i\}$  satisfying  $\text{cl}(a_t) \neq \text{cl}(a_i)$  or  $y(a_t) < y(a_j)$  or  $y(a_{j'}) < y(a_t)$ . Based on this, we can compute  $F(i, j)$  as

$$F(i, j) = \sum_{(i', j') \in J_{i,j}} (F(i', j') \cdot \pi_{i,j}^* \cdot \Pi_{i,j,i',j'}) = \pi_{i,j}^* \cdot \sum_{(i', j') \in J_{i,j}} (F(i', j') \cdot \Pi_{i,j,i',j'}). \quad (1)$$

The straightforward way to compute each  $F(i, j)$  takes  $O(n^3)$  time, which results in an  $O(n^5)$ -time algorithm for computing  $\Gamma_S$ . Indeed, the runtime of the above algorithm can be drastically improved to  $O(n^2 \log^2 n)$ , by properly using dynamic 2D range trees with some tricks. We provide some brief ideas here and defer the details to Appendix B.

A legal pair  $(i, j)$  is composed of two points,  $a_i$  and  $a_j$ . We examine all such pairs and compute the corresponding  $F(\cdot, \cdot)$  values in a fashion by first enumerating  $a_i$  from left to right and then  $a_j$  from bottom to top (when  $a_i$  is fixed). When we are about to compute  $F(i, j)$ , we need a data structure that stores  $F(i', j') \cdot \Pi_{i,j,i',j'}$  as the weight for each legal pair  $(i', j')$  and also supports range sum queries (as only those  $(i', j') \in J_{i,j}$  are of interest). A 2D range tree seems to fit here because each legal pair  $(i, j)$  can be uniquely represented by a 2D point  $(x(a_i), y(a_j))$  and we can thus associate on it the corresponding weight. On the other hand, such a range tree must be dynamic since the weights of legal pairs keep varying from time to time. We show, in Appendix B.1, how to carefully design such a data structure, and more importantly, how to use it to efficiently and correctly update those weights throughout the entire computation of  $F(i, j)$ 's. As such, we conclude the following.

**Theorem 2** *The CSD problem for  $d = 2$  can be solved in  $O(n^2 \log^2 n)$  time.*

## 2.2 Hardness results in higher dimensions

In this section, we prove the #P-hardness of the CSD problem for  $d \geq 3$ . Indeed, our hardness result is even stronger, which applies to restricted versions of the CSD problem. As introduced in Section 1, we may have two specializations of the CSD problem, in one all data points have distinct colors, in the other data points are bichromatic. We want our hardness result to cover these two specializations. To this end, we need to introduce a notion called *color pattern*.

A *partition* of a positive integer  $p$  is defined as a multi-set  $\Delta$  of positive integers whose summation is  $p$ . In a colored stochastic dataset  $\mathcal{S} = (S, \text{cl}, \pi)$ , the coloring  $\text{cl}$  naturally induces a partition of  $n = |S|$  given by the multi-set  $\{|\text{cl}^{-1}(p)| > 0 : p \in \mathbb{N}\}$ , which we denote by  $\Delta(\mathcal{S})$ . Let  $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$  be an infinite sequence where  $\Delta_p$  is a partition of  $p$ . We say  $\mathcal{P}$  is a *color pattern* if it is “polynomial-time uniform”, i.e., one can compute  $\Delta_p$  for any given  $p$  in time polynomial in  $p$ . In addition,  $\mathcal{P}$  is said to be *balanced* if  $p - \max \Delta_p = \Omega(p^c)$  for some constant  $c > 0$  (here  $\max \Delta_p$  denotes the maximum in the multi-set  $\Delta_p$ ). Then we define the CSD problem with respect to a color pattern  $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$  as the (standard) CSD problem with the restriction that the input dataset  $\mathcal{S} = (S, \text{cl}, \pi)$  must satisfy  $\Delta(\mathcal{S}) = \Delta_n$  where  $n = |S|$ .

Besides specializing the CSD problem using color pattern, we may also make assumptions for the existence probabilities of the points. An important case is that all points have the same existence probability equal to  $\frac{1}{2}$ . In this case, each of the  $2^n$  subsets of  $S$  occurs as a realization of  $\mathcal{S}$  with the same probability  $2^{-n}$ , and computing  $\Lambda_S$  (or  $\Gamma_S$ ) is equivalent to counting the subsets of  $S$  satisfying the desired properties.

Our hardness result is presented in the following theorem.

**Theorem 3** *Let  $\mathcal{P}$  be any balanced color pattern. Then the CSD problem with respect to  $\mathcal{P}$  is  $\#P$ -hard for  $d \geq 3$ . In addition, even if the existence probabilities of the points are all restricted to be  $\frac{1}{2}$ , the CSD problem with respect to  $\mathcal{P}$  remains  $\#P$ -hard for  $d \geq 7$ .*

Note that our result above implies the hardness of both the distinct-color and bichromatic specializations. The former can be seen via a balanced color pattern  $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$  with  $\Delta_p = \{1, \dots, 1\}$  (i.e., a multi-set consisting of  $p$  1's), while the latter can be seen via a balanced color pattern  $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$  with  $\Delta_p = \{\frac{p}{2}, \frac{p}{2}\}$  for even  $p$  and  $\Delta_p = \{\frac{p-1}{2}, \frac{p+1}{2}\}$  for odd  $p$ . The proof of Theorem 3 is nontrivial, so we break it into several stages.

### 2.2.1 Relation to counting independent sets

For a colored stochastic dataset  $\mathcal{S} = (S, \text{cl}, \pi)$ , define  $G_{\mathcal{S}} = (S, E_{\mathcal{S}})$  as the (undirected) graph with vertex set  $S$  and edge set  $E_{\mathcal{S}} = \{(a, b) : a, b \in S \text{ with } \text{cl}(a) \neq \text{cl}(b) \text{ and } a > b\}$ . Since the edges of  $G_{\mathcal{S}}$  one-to-one correspond to the inter-color dominances in  $\mathcal{S}$ , it is clear that a subset  $A \subseteq S$  contains no inter-color dominances iff  $A$  corresponds to an independent set of  $G_{\mathcal{S}}$ . If  $\pi(a) = \frac{1}{2}$  for all  $a \in S$ , then we immediately have the equation  $\Gamma_{\mathcal{S}} = \text{Ind}(G_{\mathcal{S}})/2^n$ , where  $\text{Ind}(G_{\mathcal{S}})$  is the number of the independent sets of  $G_{\mathcal{S}}$ . This observation intuitively tells us the hardness of the CSD problem, as independent-set counting is a well-known  $\#P$ -complete problem. Although we are still far away from proving Theorem 3 (because for a given graph  $G$  it is not clear how to construct a colored stochastic dataset  $\mathcal{S}$  such that  $G_{\mathcal{S}} \cong G$ ), it is already clear that we should reduce from some independent-set-counting problem. Regarding independent-set counting, the strongest known result is the following theorem obtained by Xia et al. [17], which will be used as the origin of our reduction.

**Theorem 4** *Counting independent sets for 3-regular planar bigraphs is  $\#P$  complete.*

For a graph  $G = (V, E)$ , we say a map  $f : V \rightarrow \mathbb{R}^d$  is a *dominance-preserving embedding* (DPE) of  $G$  to  $\mathbb{R}^d$  if it satisfies the condition that  $(u, v) \in E$  iff  $f(u) > f(v)$  or  $f(v) > f(u)$ . We define the *dimension*  $\dim(G)$  of  $G$  as the smallest number  $d$  such that there exists a DPE of  $G$  to  $\mathbb{R}^d$  (if such a number does not exist, we say  $G$  is of infinite dimension). We have seen above the relation between independent-set counting and the CSD problem with existence probabilities equal to  $\frac{1}{2}$ . Interestingly, with general existence probabilities, the CSD problem can be related to a much stronger version of independent-set counting, which we call *cardinality-sensitive independent-set counting*.

**Definition 5** *Let  $c$  be a fixed integer. The  $c$ -cardinality-sensitive independent-set counting ( $c$ -CSISC) problem is defined as follows. The input consists of a graph  $G = (V, E)$  and a  $c$ -tuple  $\Phi = (V_1, \dots, V_c)$  of disjoint subsets of  $V$ . The task of the problem is to output, for every  $c$ -tuple  $(n_1, \dots, n_c)$  of integers where  $0 \leq n_i \leq |V_i|$ , the number of the independent sets  $I \subseteq V$  of  $G$  satisfying  $|I \cap V_i| = n_i$  for all  $i \in \{1, \dots, c\}$ . We denote the desired output by  $\text{Ind}_{\Phi}(G)$ , which can be represented by a sequence of  $\prod_{i=1}^c (|V_i| + 1)$  integers. Note that the 0-CSISC problem is just the conventional independent-set counting.*

**Lemma 6** *Given any graph  $G = (V, E)$  with a DPE  $f : V \rightarrow \mathbb{R}^d$  and a  $c$ -tuple  $\Phi = (V_1, \dots, V_c)$  of disjoint subsets of  $V$ , one can construct in polynomial time a colored stochastic dataset  $\mathcal{S} = (S, \text{cl}, \pi)$  in  $\mathbb{R}^d$  with  $\text{cl}$  injective such that (1)  $G_{\mathcal{S}} \cong G$  and (2)  $\text{Ind}_{\Phi}(G)$  can be computed in polynomial time if  $\Gamma_{\mathcal{S}}$  is provided. In particular, the  $c$ -CSISC problem for a class  $\mathcal{G}$  of graphs is polynomial-time reducible to the CSD problem in  $\mathbb{R}^d$ , provided an oracle that computes for any graph in  $\mathcal{G}$  a DPE to  $\mathbb{R}^d$ .*

Indeed, Lemma 6 has a more general version which reveals the hardness of (general) stochastic geometric problems, see Appendix C for details. Another ingredient to be used in the proof of Theorem 3 is a lemma regarding color pattern.

**Lemma 7** *Let  $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$  be a balanced color pattern. Given a colored stochastic dataset  $\mathcal{S} = (S, \text{cl}, \pi)$  in  $\mathbb{R}^d$  with  $\text{cl}$  injective, if  $G_{\mathcal{S}}$  is a bipartite graph, then one can construct in polynomial time another colored stochastic dataset  $\mathcal{S}' = (S', \text{cl}', \pi')$  in  $\mathbb{R}^d$  satisfying (1)  $\Gamma_{\mathcal{S}'} = \Gamma_{\mathcal{S}}$ , (2)  $S \subseteq S'$ , (3)  $\pi'(a) = \frac{1}{2}$  for any  $a \in S \setminus S$ , (4)  $\langle S' \rangle$  is an instance of the CSD problem with respect to  $\mathcal{P}$ .*

### 2.2.2 #P-hardness for $d \geq 3$

In this section, we prove the first statement of Theorem 3, by providing a reduction from the independent-set counting problem for 3-regular planar bipartite graphs. Let  $G = (V \cup V', E)$  be a 3-regular planar bipartite graph. Suppose  $|V| = |V'| = n$  (note that we must have  $|V| = |V'|$  for  $G$  is 3-regular), and then  $|E| = 3n$ . Instead of working on  $G$  directly, we first pass to a new graph, which seems to have a lower dimension. Set  $\lambda = 100n^2$ . We define  $G^*$  as the graph obtained from  $G$  by inserting  $2\lambda$  new vertices to each edge of  $G$ , i.e., replacing each edge of  $G$  with a chain of  $2\lambda$  new vertices (see Figure 2). With an abuse of notation,  $V$  and  $V'$  are also used to denote the corresponding subsets of the vertices of  $G^*$ . Note that  $G^*$  is also bipartite, in which  $V$  and  $V'$  belong to different parts. We use  $U$  (resp.,  $U'$ ) to denote the set of the inserted vertices of  $G^*$  which belong to the same part as  $V$  (resp.,  $V'$ ). Then the two parts of  $G^*$  are  $V \cup U$  and  $V' \cup U'$ . For each edge  $e \in E$  of  $G$ , we denote by  $U_e$  (resp.  $U'_e$ ) the set of the  $\lambda$  vertices in  $U$  (resp.,  $U'$ ) which are inserted to the edge  $e$ .

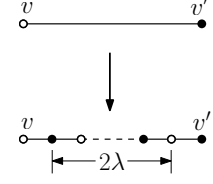


Figure 2: Inserting new vertices to each edge of  $G$ .

It is not surprising that the independent sets of  $G$  are strongly related to those of  $G^*$ . Indeed, as we will show, counting independent sets for  $G$  can be done by solving the 4-CSISC instance  $\langle G^*, (V, V', U, U') \rangle$ . Define  $Ind_{p,p'}$  as the number of the independent sets  $I$  of  $G$  such that  $|I \cap V| = p$ ,  $|I \cap V'| = p'$ . Also, define  $Ind_{p,p',q,q'}^*$  as the number of the independent sets  $I^*$  of  $G^*$  such that  $|I^* \cap V| = p$ ,  $|I^* \cap V'| = p'$ ,  $|I^* \cap U| = q$ ,  $|I^* \cap U'| = q'$ .

**Lemma 8** *For any  $p, p' \in \{0, \dots, n\}$ , we have  $Ind_{p,p'} = Ind_{p,p',3\lambda p,3\lambda n-3\lambda p}^*$ . In particular,*

$$Ind(G) = \sum_{i=0}^n \sum_{j=0}^n Ind_{i,j} = \sum_{i=0}^n \sum_{j=0}^n Ind_{i,j,3\lambda i,3\lambda n-3\lambda i}^*.$$

Now it suffices to reduce the 4-CSISC instance  $\langle G^*, (V, V', U, U') \rangle$  to an instance  $\langle S \rangle$  of the CSD problem in  $\mathbb{R}^3$  with respect to a given balanced color pattern  $\mathcal{P}$ . Due to Lemma 6 and 7, the only thing we need for the reduction is a DPE of  $G^*$  to  $\mathbb{R}^3$ . Therefore, our next step is to show  $\dim(G^*) \leq 3$  and construct explicitly a DPE of  $G^*$  to  $\mathbb{R}^3$  (in polynomial time), which is the most non-obvious part of the proof.

Recall that the two parts of  $G^*$  are  $V \cup U$  and  $V' \cup U'$ . The DPE that we are going to construct makes the image of each vertex in  $V' \cup U'$  dominates the images of its adjacent vertices in  $V \cup U$ . We first consider the embedding for the part  $V \cup U$ . Our basic idea is to map the vertices in  $V \cup U$  to the plane  $H : x + y + z = 0$  in  $\mathbb{R}^3$ . Note that by doing this we automatically prevent their images from dominating each other. However, the locations of (the images of) these vertices on  $H$  should be carefully chosen so that later we are able to further embed the part  $V' \cup U'$  (to  $\mathbb{R}^3$ ) to form a DPE. Basically, we map  $V \cup U$  to  $H$  through two steps. In the first step, the vertices in  $V \cup U$  are mapped to  $\mathbb{R}^2$  via a map  $\varphi : V \cup U \rightarrow \mathbb{R}^2$  to be constructed. Then in the second step, we properly project  $\mathbb{R}^2$  onto  $H$  via another map  $\psi : \mathbb{R}^2 \rightarrow H$ . By composing  $\psi$  and  $\varphi$ , we obtain the desired map  $\psi \circ \varphi : V \cup U \rightarrow H$ , which gives us the embedding for  $V \cup U$ .

To construct  $\varphi$ , we need a notion about graph drawing. Let  $K = (\mathbb{Z} \times \mathbb{R}) \cup (\mathbb{R} \times \mathbb{Z}) \subset \mathbb{R}^2$  be the grid. An *orthogonal grid drawing* (OGD) of a (planar) graph is a planar drawing with image in the grid  $K$  such that the vertices are mapped to the grid points  $\mathbb{Z}^2$ . Note that an OGD draws the edges of the graph as (non-intersecting) orthogonal curves in  $\mathbb{R}^2$  consisting of unit-length horizontal/vertical segments each of which connects two adjacent grid points (see Figure 3). We will apply the following result from [16].

**Theorem 9** *For any  $t$ -vertex planar graph of (maximum) degree 3, one can compute in polynomial time an OGD with image in  $K \cap Q_{3t}$  where  $Q_i$  denotes the square  $[1, i] \times [1, i] \subset \mathbb{R}^2$ .*

Consider the original 3-regular planar bipartite graph  $G = (V \cup V', E)$ . By applying the above theorem, we can find an OGD  $g$  for  $G$  with image in  $K \cap Q_{6n}$ . For each vertex  $v \in V \cup V'$  of  $G$ , we denote by  $g(v)$  the image of  $v$  in  $\mathbb{R}^2$  under the OGD  $g$ . Also, for each edge  $e = (v, v') \in E$  of  $G$ , we denote by  $g(e)$  the image of  $e$  under  $g$ , which is an orthogonal curve in  $\mathbb{R}^2$  connecting  $g(v)$  and  $g(v')$ . With the OGD  $g$  in hand, we construct the map  $\varphi$  as follows. For all  $v \in V$ , we simply define  $\varphi(v) = g(v)$ . To determine  $\varphi(u)$  for  $u \in U$ ,

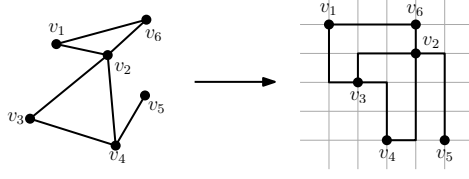


Figure 3: An orthogonal grid drawing.

we consider the vertices in  $U_e$  for each edge  $e \in E$  of  $G$  separately. Suppose  $e = (v, v')$  and  $U_e = \{u_1, \dots, u_\lambda\}$  where  $u_1, \dots, u_\lambda$  are sorted in the order they appear on  $e$  (from  $v$  to  $v'$ ). Consider the curve  $g(e)$ . Since  $g$  is an OGD,  $g(e)$  must consist of unit-length horizontal/vertical segments (each of which connects two grid points). The total number  $m$  of these unit segments is upper bounded by  $(6n)^2$  as  $g(e) \subset K \cap Q_{6n}$ . Now we pick a set  $P_e$  of  $\lambda$  (distinct) points on  $g(e)$  as follows.

- The  $m - 1$  grid points in the interior of  $g(e)$  are included in  $P_e$  (see Figure 4a).
- On each unit vertical segment of  $g(e)$ , we pick the point with distance 0.3 from the bottom endpoint and include it to  $P_e$  (see Figure 4b).
- On the unit segment of  $g(e)$  adjacent to  $g(v')$ , we pick the point with distance 0.01 from  $g(v')$  and include it to  $P_e$  (see Figure 4c).
- Note that the number of the above three types of points is at most  $2m \leq 72n^2 < \lambda$ . To make  $|P_e| = \lambda$ , we then arbitrarily pick more (distinct) points on  $g(e)$  which have distances at least 0.4 to any grid point, and add them to  $P_e$ .

Suppose  $P_e = \{r_1, \dots, r_\lambda\}$  where  $r_1, \dots, r_\lambda$  are sorted in the order they appear on the curve  $g(e)$  (from  $g(v)$  to  $g(v')$ ). We then define  $\varphi(u_i) = r_i$ . We do the same thing for every edge  $e \in E$  of  $G$ . In this way, we determine  $\varphi(u)$  for all  $u \in U$  and complete defining the map  $\varphi$ . The next step, as mentioned before, is to project  $\mathbb{R}^2$  onto  $H$ . The projection map  $\psi : \mathbb{R}^2 \rightarrow H$  is defined as  $\psi : (x, y) \mapsto (x + y, y - x, -2y)$ . Then the composition  $\psi \circ \varphi : V \cup U \rightarrow H$  gives us the first part of our DPE. The remaining task is to embed the part

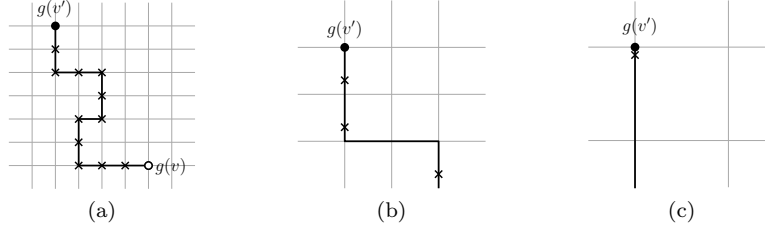


Figure 4: The construction of  $P_e$ .

$V' \cup U'$  to  $\mathbb{R}^3$ , which completes the construction of our DPE. We must guarantee that the image of each vertex  $w' \in V' \cup U'$  dominates and only dominates the images of the vertices in  $V \cup U$  adjacent to  $w'$ . To achieve this, we first establish an important property of the map  $\psi \circ \varphi : V \cup U \rightarrow H$  constructed above. For a finite set  $A$  of points in  $\mathbb{R}^d$ , we define a point  $\underline{\max}(A) \in \mathbb{R}^d$  as the *coordinate-wise* maximum of  $A$ , i.e., the  $i$ -th coordinate of  $\underline{\max}(A)$  is the maximum of the  $i$ -th coordinates of all points in  $A$ , for all  $i \in \{1, \dots, d\}$ .

**Lemma 10** *For each vertex  $w' \in V' \cup U'$ , let  $\text{Adj}_{w'} \subseteq V \cup U$  be the set of the vertices adjacent to  $w'$  in  $G^*$ , and  $A_{w'} = (\psi \circ \varphi)(\text{Adj}_{w'}) \subset \mathbb{R}^3$  be the set of the corresponding images under  $\psi \circ \varphi$ . Then for any  $w \in V \cup U$  and  $w' \in V' \cup U'$ , the point  $\underline{\max}(A_{w'}) \in \mathbb{R}^3$  dominates  $(\psi \circ \varphi)(w)$  iff  $w \in \text{Adj}_{w'}$ .*

Once the above property is revealed, the construction of the map  $V' \cup U' \rightarrow \mathbb{R}^3$  is quite simple: we just map each vertex  $w' \in V' \cup U'$  to the point  $\underline{\max}(A_{w'}) \in \mathbb{R}^3$ . Now we complete constructing the embedding of  $G^*$  to  $\mathbb{R}^3$ , and need to verify it is truly a DPE. Lemma 10 already guarantees that the image of each  $w' \in V' \cup U'$  dominates (the images of) the vertices in  $\text{Adj}_{w'}$  (i.e., the vertices in  $V \cup U$  that are adjacent to  $w'$ ) but does not dominate (the images of) any other vertices in  $V \cup U$ . So it suffices to show that the

images of the vertices in  $V' \cup U'$  do not dominate each other. Let  $w'_1, w'_2 \in V' \cup U'$  be two distinct vertices, and assume that  $\max(A_{w'_1})$  dominates  $\max(A_{w'_2})$ . Then we must have  $\max(A_{w'_1})$  dominates the points in  $A_{w'_2}$ . By Lemma 10, this implies that  $\text{Adj}_{w'_2} \subseteq \text{Adj}_{w'_1}$ . However, as one can easily see from the structure of  $G^*$ , it never happens that  $\text{Adj}_{w'_2} \subseteq \text{Adj}_{w'_1}$  unless  $w'_1 = w'_2$ . Thus, we conclude that the map constructed is a DPE of  $G^*$  to  $\mathbb{R}^3$ . With the DPE in hand, by applying Lemma 6 and 7, the first statement of Theorem 3 is readily proved.

### 2.2.3 #P-hardness for $d \geq 7$ with half existence probabilities

In this section, we prove the second statement of Theorem 3. When the existence probabilities are restricted to be  $\frac{1}{2}$ , we are no longer able to apply the tricks used in the previous section, as the reduction from the CSISC problem (Lemma 6) cannot be done under such a restriction. This is the reason for why we have to “loosen” the dimension to 7 in this case.

As we have seen, for a colored stochastic dataset  $\mathcal{S} = (S, \text{cl}, \pi)$  with  $\pi(a) = \frac{1}{2}$  for all  $a \in S$ , computing  $\Gamma_{\mathcal{S}}$  is totally equivalent to counting independent sets for  $G_{\mathcal{S}}$ . Therefore, we complete the proof by establishing a more direct reduction from independent-set counting for 3-regular planar bipartite graphs, which constructs directly a DPE of the input graph to  $\mathbb{R}^7$ . However, it is non-obvious that any 3-regular planar bipartite graph  $G$  has dimension at most 7 and how to construct a DPE of  $G$  to  $\mathbb{R}^7$  in polynomial time. To prove this, we introduce a new technique based on graph coloring. Indeed, we consider a more general case in which the graph  $G$  is an arbitrary bipartite graph. The graph coloring to be used is slightly different from the conventional notion, which we call *halfcoloring*. Let  $G = (V \cup V', E)$  be a bipartite graph. For any two distinct vertices  $u, v \in V$ , we define  $u \sim v$  if there exists a vertex in  $V'$  adjacent to both  $u$  and  $v$ .

**Definition 11** *A  $k$ -halfcoloring of  $G$  on  $V$  is a map  $h : V \rightarrow \{1, \dots, k\}$ . The halfcoloring  $h$  is said to be discrete if  $h(u) \neq h(v)$  for any  $u, v \in V$  with  $u \sim v$ , to be semi-discrete if it satisfies the condition that for any distinct  $u, v, w \in V$  with  $u \sim v$  and  $v \sim w$ ,  $h(u), h(v), h(w)$  are not all the same. Symmetrically, we may also define halfcoloring on  $V'$ .*

We may relate halfcoloring to the conventional graph coloring as follows. Define  $G' = (V, E')$  with  $E' = \{(u, v) : u \sim v \text{ in } G\}$ . Clearly, a discrete  $k$ -halfcoloring of  $G$  on  $V$  corresponds to a (conventional)  $k$ -coloring of  $G'$  satisfying that no two adjacent vertices share the same color, i.e., the subgraph of  $G'$  induced by each color form an independent set of  $G'$ . Similarly, a semi-discrete  $k$ -halfcoloring of  $G$  on  $V$  corresponds to a  $k$ -coloring of  $G'$  satisfying that the subgraph of  $G'$  induced by each color consists of connected components of sizes at most 2. If  $h$  is a  $k$ -halfcoloring of  $G$  on  $V$ , then for each  $v' \in V'$  we denote by  $\chi_h(v')$  the number of the colors “adjacent” to  $v'$  (the color  $i$  is said to be adjacent to  $v'$  if there is a vertex  $v \in V$  adjacent to  $v'$  with  $h(v) = i$ ). Our technical result is the following theorem, which establishes a relation between halfcoloring and graph dimension.

**Theorem 12** *Let  $G = (V \cup V', E)$  be a bipartite graph. If there exists a semi-discrete  $k$ -halfcoloring  $h : V \rightarrow \{1, \dots, k\}$  of  $G$  (on  $V$ ), then  $\dim(G) \leq 2k$ . In addition, if  $\chi_h(v') < k$  for all  $v' \in V'$ , then  $\dim(G) \leq 2k - 1$ . Furthermore, with  $h$  in hand, one can compute in polynomial time a DPE of  $G$  to  $\mathbb{R}^{2k}$ , or  $\mathbb{R}^{2k-1}$  in the latter case.*

We then apply the halfcoloring technique to show that  $\dim(G) \leq 7$  for any 3-regular planar bipartite graph  $G$ , which will give us a proof for the second statement of Theorem 3. To achieve this, the only missing piece is the following observation.

**Lemma 13** *Every 3-regular planar bipartite graph has a discrete 4-halfcoloring, which can be computed in polynomial time.*

Now it is quite straightforward to prove the second statement of Theorem 3. Let  $G$  be a 3-regular planar bipartite graph. By combining Theorem 12 and Lemma 13, we can compute a DPE of  $G$  to  $\mathbb{R}^7$  in polynomial time. By taking the images of the vertices of  $G$  under the DPE, we obtain a set  $S$  of points in  $\mathbb{R}^7$ . Using the point set  $S$ , we further construct a colored stochastic dataset  $\mathcal{S} = (S, \text{cl}, \pi)$  by choosing an injection



$\text{cl} : S \rightarrow \mathbb{N}$  and defining  $\pi(a) = \frac{1}{2}$  for any  $a \in S$ . It is clear that  $G_S \cong G$  and thus  $\text{Ind}(G) = 2^{|S|} \Gamma_S$ . Then by applying Lemma 7, we can compute another colored stochastic dataset  $\mathcal{S}' = (S', \text{cl}', \pi')$  such that  $\Gamma_{\mathcal{S}'} = \Gamma_S$  and  $\pi'(a) = \frac{1}{2}$  for any  $a \in S'$ , and more importantly,  $\langle \mathcal{S}' \rangle$  is an instance of the CSD problem with respect to  $\mathcal{P}$ . With this reduction, the second statement of Theorem 3 is proved. Our conclusion that any 3-regular planar bipartite graph has dimension at most 7 is of independent interest, which results in an implication in order dimension theory [15] (see Appendix D).

### 2.3 A simple FPRAS

In this section, we describe a simple FPRAS (i.e., fully polynomial-time randomized approximation scheme) for approximating  $\Lambda_S$  in any dimension. Recall that a FPRAS is a randomized algorithm which takes the input of the problem with an additional parameter  $\varepsilon > 0$ , and computes an  $\varepsilon$ -approximation of the answer in polynomial (in both the size of the problem and  $1/\varepsilon$ ) time with high probability (say at least  $2/3$ ).

A natural idea to design a FPRAS for approximating  $\Lambda_S$  is to randomly generate a large number of realizations of  $\mathcal{S}$ , and estimate  $\Lambda_S$  using the proportion of the number of the realizations containing inter-color dominances to the total number of the realizations. However, since we are only allowed to generate polynomial number of realizations, this method does not guarantee to produce an  $\varepsilon$ -approximation of  $\Lambda_S$  with high probability. For instance, if  $\Lambda_S = 2^{-n}$ , then the estimation of  $\Lambda_S$  obtained by generating polynomial number of realizations would be 0 with probability almost 1 (as one can easily verify using union bound). Interestingly, by slightly making some changes to this simple method, we can truly obtain a FPRAS for computing  $\Lambda_S$ .

Our FPRAS works as follows. Suppose the points  $a_1, \dots, a_n$  are already sorted by their existence probabilities from large to small, i.e.,  $\pi(a_1) \geq \dots \geq \pi(a_n)$ . Instead of estimating  $\Lambda_S$  directly, what we do is to estimate a set of conditional probabilities and use them to compute an estimation of  $\Lambda_S$ . For any  $i, j \in \{1, \dots, n\}$  with  $i < j$ , we define  $E_{i,j}$  as the event that a realization  $R$  of  $\mathcal{S}$  includes  $a_i, a_j$  and any other points in  $R$  have indices smaller than  $i$ . Then we immediately have

$$\Lambda_S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \Pr[E_{i,j}] \cdot \text{Cond}_{i,j}, \quad (2)$$

where  $\text{Cond}_{i,j}$  is the conditional probability that a realization of  $\mathcal{S}$  contains inter-color dominances under the condition that  $E_{i,j}$  happens. The probabilities  $\Pr[E_{i,j}]$  can be straightforwardly computed. But we are not able to exactly compute  $\text{Cond}_{i,j}$  in polynomial time, so we try to estimate them by randomly generating realizations. For  $p \in \{0, 1, \dots, n\}$ , set  $S_p = \{a_1, \dots, a_p\}$ , and we use  $\mathcal{S}_p$  to denote the sub-dataset of  $\mathcal{S}$  with point set  $S_p \subseteq S$ . We randomly generate  $N = 10n^5/\varepsilon^2$  realizations of  $\mathcal{S}_p$  for each  $p \in \{0, 1, \dots, n\}$ . Let  $R_{p,q}$  be the  $q$ -th realization of  $\mathcal{S}_p$ . Naturally, we compute an estimation  $\text{Est}_{i,j}$  for each  $\text{Cond}_{i,j}$  as

$$\text{Est}_{i,j} = \sum_{k=1}^N \frac{\sigma(R_{i-1,k} \cup \{a_i, a_j\})}{N},$$

where  $\sigma(R) = 1$  if  $R$  contains inter-color dominances and  $\sigma(R) = 0$  otherwise. Then we can apply Equation 2 to compute an estimation  $\Lambda$  of  $\Lambda_S$ , simply by replacing each  $\text{Cond}_{i,j}$  with its estimation  $\text{Est}_{i,j}$ . It is quite surprising that  $\Lambda$  is, with high probability, an  $\varepsilon$ -approximation of  $\Lambda_S$  (note that each  $\text{Est}_{i,j}$  is not necessarily an  $\varepsilon$ -approximation of  $\text{Cond}_{i,j}$  with high probability). The following theorem completes the discussion.

**Theorem 14** *We have  $(1 - \varepsilon)\Lambda_S < \Lambda < (1 + \varepsilon)\Lambda_S$  with probability at least  $2/3$ .*

## 3 The free-basis colored stochastic dominance problem

Let  $\mathcal{S} = (S, \text{cl}, \pi)$  be a colored stochastic dataset in  $\mathbb{R}^d$  with  $S = \{a_1, \dots, a_n\}$ . By naturally generalizing the conventional dominance relation, one can define dominance relation with respect to a specific orthogonal basis

of  $\mathbb{R}^d$ . For two distinct points  $p, q \in \mathbb{R}^d$  and an orthogonal basis  $B = (\mathbf{b}_1, \dots, \mathbf{b}_d)$  of  $\mathbb{R}^d$ , we say  $p$  *dominates*  $q$  *with respect to*  $B$  (denoted by  $p >_B q$ ) if  $\langle \mathbf{b}_i, p \rangle \geq \langle \mathbf{b}_i, q \rangle$  for every  $i \in \{1, \dots, d\}$ , where  $\langle \cdot, \cdot \rangle$  is the inner product. With this generalized definition, the conventional dominance relation is just dominance relation with respect to the standard basis  $E = (\mathbf{e}_1, \dots, \mathbf{e}_d)$  of  $\mathbb{R}^d$ . Define  $\Lambda_S^*$  as the probability that a realization of  $S$  contains inter-color dominances with respect to any orthogonal basis of  $\mathbb{R}^d$ . Set  $\Gamma_S^* = 1 - \Lambda_S^*$ , which is the probability that a realization of  $S$  contains no inter-color dominances with respect to some orthogonal basis of  $\mathbb{R}^d$ . The goal of the FBCSD problem is to compute  $\Lambda_S^*$  (or  $\Gamma_S^*$ ).

### 3.1 Reduction from the CSD problem

In this section, we show that the (standard) CSD problem in  $\mathbb{R}^d$  is polynomial-time reducible to the FBCSD problem in the same dimension, which implies the latter is  $\#P$ -hard for  $d \geq 3$ .

Given a colored stochastic dataset  $S = (S, \text{cl}, \pi)$  in  $\mathbb{R}^d$  as an instance of the CSD problem, our reduction tries to construct another colored stochastic dataset  $S' = (S', \text{cl}', \pi')$  in  $\mathbb{R}^d$  such that  $\Lambda_{S'}^* = \Lambda_S$ . The intuition of our reduction is the following. First, consider the given colored stochastic dataset  $S$ . Clearly, we have  $\Lambda_S^* \leq \Lambda_S$ , as every realization of  $S$  counted in  $\Lambda_S^*$  is also counted in  $\Lambda_S$ . The reason for why  $\Lambda_S^*$  may be smaller than  $\Lambda_S$  is that perhaps some realization contains inter-color dominances with respect to the standard basis  $E$  of  $\mathbb{R}^d$  but does not contain inter-color dominances with respect to some other basis. To handle this, our basic idea is to add a set  $\Psi$  of (colored) auxiliary points with existence probabilities 1 to  $S$ , that is, we want  $S' = S \cup \Psi$  with  $\pi'(b) = 1$  for all  $b \in \Psi$  (and  $\pi'(a) = \pi(a)$ ,  $\text{cl}'(a) = \text{cl}(a)$  for all  $a \in S$ ). The goal of adding these auxiliary points is to guarantee that a subset  $A \subseteq S$  contains inter-color dominances with respect to the standard basis  $E$  iff  $A \cup \Psi \subseteq S'$  contains inter-color dominances with respect to any orthogonal basis. Note that as long as  $\Psi$  has this property, it obviously holds that  $\Lambda_{S'}^* = \Lambda_S$ . Therefore, the critical part of our reduction is to construct such a set  $\Psi$  with the desired property. We achieve this through several steps.

First of all, we need to make the point set  $S$  “regular”. Formally, we say a (finite) point set  $X \subset \mathbb{R}^d$  is *regular* if  $X \subset \{1, 2, \dots, |X|\}^d$  and any two distinct points  $x, x' \in X$  have distinct coordinates in all dimensions. It is easy to see that one can always “regularize” a point set without changing the dominance relation (with respect to  $E$ ) among the points.

**Lemma 15** *Given a set  $S = \{a_1, \dots, a_n\} \subset \mathbb{R}^d$  of distinct points, one can construct in  $O(n \log n)$  time a regular set  $S_{\text{new}} = \{\hat{a}_1, \dots, \hat{a}_n\} \subset \mathbb{R}^d$  such that  $\hat{a}_i >_E \hat{a}_j$  iff  $a_i >_E a_j$ .*

Now we may assume  $S$  is regular. To construct  $\Psi$ , we need to introduce some notions.

**Definition 16** *Let  $B = (\mathbf{b}_1, \dots, \mathbf{b}_d)$  be an orthogonal basis of  $\mathbb{R}^d$ . We define the cone  $C_B$  of  $B$  as*

$$C_B = \left\{ \sum_{i=1}^d \beta_i \mathbf{b}_i : \beta_1, \dots, \beta_d \geq 0 \right\} \cup \left\{ \sum_{i=1}^d \beta_i \mathbf{b}_i : \beta_1, \dots, \beta_d \leq 0 \right\} \subset \mathbb{R}^d.$$

*Also, we define the projective cone  $PC_B \subset \mathbb{P}^{d-1}$  as the image of  $C_B \setminus \{\mathbf{0}\}$  in  $\mathbb{P}^{d-1}$  under the obvious quotient map  $\mathbb{R}^d \setminus \{\mathbf{0}\} \rightarrow \mathbb{P}^{d-1}$ .*

For a point  $x \in \mathbb{R}^d$  with  $x \neq \mathbf{0}$ , we denote by  $\bar{x}$  its image in  $\mathbb{P}^{d-1}$  under the quotient map  $\mathbb{R}^d \setminus \{\mathbf{0}\} \rightarrow \mathbb{P}^{d-1}$ . The notion of (projective) cone defined above gives us another way to view dominance relations with respect to an orthogonal basis. Consider two distinct points  $p, q \in \mathbb{R}^d$ , and an orthogonal basis  $B$  of  $\mathbb{R}^d$ . It is easy to see that  $p, q$  form a dominance with respect to  $B$  (i.e.,  $p >_B q$  or  $q >_B p$ ) iff  $p - q \in C_B$ , or equivalently,  $\overline{p - q} \in PC_B$ . Another notion we need is a metric on any projective space  $\mathbb{P}^k$ .

**Definition 17** *For two points  $l, l' \in \mathbb{P}^k$ , we define  $\text{ang}(l, l') \in [0, \frac{\pi}{2}]$  to be the angle between  $l$  and  $l'$  as lines in  $\mathbb{R}^{k+1}$  through the origin (there are two supplementary angles, take the smaller one which is in  $[0, \frac{\pi}{2}]$ ). It is easy to see that  $\text{ang}(\cdot, \cdot)$  defines a metric on  $\mathbb{P}^k$ .*

The following lemmas establish some geometric properties of the projective cone and the ang-metric, which will be helpful for constructing  $\Psi$ .

**Lemma 18** Let  $B$  be an orthogonal basis of  $\mathbb{R}^d$ , and  $l$  be a point in  $\mathbb{P}^{d-1}$ . If  $l \notin PC_B$ , then there exists  $x \in PC_B$  perpendicular to  $l$ , i.e.,  $\text{ang}(l, x) = \frac{\pi}{2}$ .

**Lemma 19** For any orthogonal basis  $B$  of  $\mathbb{R}^d$ , any point  $x \in PC_B$ , and any real number  $\varepsilon \in (0, \frac{\pi}{2}]$ , there exists  $y \in PC_B$  with  $\text{ang}(x, y) < \varepsilon$  such that the  $\frac{\varepsilon}{3\sqrt{d}}$ -ball at  $y$ , i.e., the set  $\{z \in \mathbb{P}^{d-1} : \text{ang}(z, y) \leq \frac{\varepsilon}{3\sqrt{d}}\}$ , is contained in  $PC_B$ .

**Lemma 20** Let  $l$  be a point in  $\mathbb{P}^{d-1}$  and  $\varepsilon \geq \xi > 0$  be two real numbers. Then one can compute  $m = O(\varepsilon/\xi^{d-1})$  points  $l_1, \dots, l_m \in \mathbb{P}^{d-1}$  in  $O(m)$  time such that (1)  $\text{ang}(l, l_i) > \frac{\pi}{2} - \varepsilon$  for all  $i \in \{1, \dots, m\}$  and (2) for any  $y \in \mathbb{P}^{d-1}$  with  $\text{ang}(l, y) > \frac{\pi}{2} - \varepsilon$ , there exists some  $l_i$  satisfying  $\text{ang}(l_i, y) < \xi$ .

With the above lemmas in hand, we now describe the construction of  $\Psi$ . We look at all pairs  $(a, a')$  of points in  $S$  such that  $\text{cl}(a) \neq \text{cl}(a')$  and  $a >_E a'$ . For each such pair  $(a, a')$ , we do the following. Set  $l = \overline{a - a'} \in \mathbb{P}^{d-1}$ ,  $\varepsilon = \arcsin(\frac{1}{\sqrt{dn}})$ , and  $\xi = \frac{\varepsilon}{3\sqrt{d}}$ . By applying Lemma 20 with  $l, \varepsilon, \xi$ , we compute  $m = O(\varepsilon/\xi^{d-1}) = O(n^{d-2})$  points  $l_1, \dots, l_m \in \mathbb{P}^{d-1}$  satisfying the conditions (1) and (2) in the lemma. In addition, we observe the following.

- $l_i \notin PC_E$  for all  $i \in \{1, \dots, m\}$ .
- For any orthogonal basis  $B$  of  $\mathbb{R}^d$ , if  $l \notin PC_B$ , then there exists some  $l_i \in PC_B$ .

To see the first observation, recall that  $S$  is already regular. Since  $a >_E a'$  and  $S$  is regular,  $l$  can be represented by homogeneous coordinates  $[\alpha_1 : \dots : \alpha_d]$  with  $\alpha_1, \dots, \alpha_d \in \{1, \dots, n-1\}$ . Based on this, one can easily verify that  $\text{ang}(l, l') < \arccos(\frac{1}{\sqrt{dn}})$  for any  $l' \in PC_E$ . But we have  $\text{ang}(l, l_i) > \frac{\pi}{2} - \varepsilon = \arccos(\frac{1}{\sqrt{dn}})$  by Lemma 20. Thus,  $l_i \notin PC_E$ . To see the second observation, let  $B$  be an orthogonal basis of  $\mathbb{R}^d$  with  $l \notin PC_B$ . By Lemma 18, there exists  $x \in PC_B$  with  $\text{ang}(l, x) = \frac{\pi}{2}$ . Then by Lemma 19, there exists  $y \in PC_B$  such that  $\text{ang}(x, y) < \varepsilon$  and the  $\frac{\varepsilon}{3\sqrt{d}}$ -ball at  $y$  is contained in  $PC_B$ . Since  $\text{ang}(l, x) = \frac{\pi}{2}$  and  $\text{ang}(x, y) < \varepsilon$ , we have  $\text{ang}(l, y) > \frac{\pi}{2} - \varepsilon$ . Therefore, according to the condition (2) in Lemma 20, there must exist some  $l_i$  such that  $\text{ang}(l_i, y) < \xi$ . Recall that  $\xi = \frac{\varepsilon}{3\sqrt{d}}$ , so  $l_i$  is in the  $\frac{\varepsilon}{3\sqrt{d}}$ -ball at  $y$  and hence in  $PC_B$ . These two observations will be used later to verify that  $\Psi$  satisfies the desired property. Now we continue to discuss the construction of  $\Psi$ . We have computed  $m$  points  $l_1, \dots, l_m \in \mathbb{P}^{d-1}$  for a specific pair  $(a, a')$ . We do the same thing for all pairs  $(a, a')$  of points in  $S$  with  $\text{cl}(a) \neq \text{cl}(a')$  and  $a >_E a'$ . After this, we obtain  $M = O(n^2m) = O(n^d)$  points in  $\mathbb{P}^{d-1}$  (with an abuse of notation, we denote them by  $l_1, \dots, l_M$ ). The set  $\Psi$  we construct consists of  $2M$  points  $b_1, \dots, b_M, b'_1, \dots, b'_M \in \mathbb{R}^d$  where  $b_i, b'_i$  correspond to  $l_i$  for  $i \in \{1, \dots, M\}$ . We set the coordinates of each  $b_i$  in  $\mathbb{R}^d$  to be  $(-i, \dots, -i, n+i)$ . Then we choose location for each  $b'_i$  in  $\mathbb{R}^d$  such that  $\|b'_i - b_i\|_2 < 0.1$  and  $\overline{b'_i - b_i} = l_i$  (there are infinitely many choices, we arbitrarily pick one of them). Finally, we need to define the coloring of the points in  $\Psi$ , i.e.,  $\text{cl}'(b)$  for all  $b \in \Psi$ . We arbitrarily color the points in  $\Psi$  under the only restriction that  $b_i$  and  $b'_i$  must have different colors, i.e.,  $\text{cl}'(b_i) \neq \text{cl}'(b'_i)$ , for all  $i \in \{1, \dots, M\}$ . It suffices to verify the property that  $A \subseteq S$  contains inter-color dominances with respect to  $E$  iff  $A \cup \Psi \subseteq S'$  contains inter-color dominances with respect to any orthogonal basis. To see the “if” part, let  $A \subseteq S$  be a subset such that  $A \cup \Psi \subseteq S'$  contains inter-color dominances with respect to any orthogonal basis. Since  $S \subset [1, n]^d$  (as  $S$  is regular) and  $l_i \notin PC_E$  for all  $i \in \{1, \dots, M\}$  (as observed above), the points in  $\Psi$  do not dominate each other and do not form dominance with any points in  $S$ , with respect to  $E$ . But by assumption,  $A \cup \Psi$  contains inter-color dominances with respect to  $E$ . So the inter-color dominances must be formed by the points in  $A$ , i.e.,  $A$  contains inter-color dominances with respect to  $E$ . To see the “only if” part, let  $A \subseteq S$  be a subset containing inter-color dominances with respect to  $E$ . Suppose  $a, a' \in A$  are two points such that  $\text{cl}(a) \neq \text{cl}(a')$  and  $a >_E a'$ . Consider an orthogonal basis  $B$  of  $\mathbb{R}^d$ , and we must show that  $A \cup \Psi$  contains inter-color dominances with respect to  $B$ . If  $a >_B a'$  or  $a' >_B a$ , then we are done. Otherwise, recall that we have  $m$  points in  $\{l_1, \dots, l_M\}$  which are chosen for the pair  $(a, a')$  (assume they are  $l_1, \dots, l_m$  without loss of generality). By our observation above, one of these  $m$  points must be in  $PC_B$ , say  $l_1 \in PC_B$ . Then the two points  $b_1, b'_1 \in \Psi$  form an inter-color dominance with respect to  $B$ .

By the above construction, we obtain a colored stochastic dataset  $\mathcal{S}' = (S \cup \Psi, \text{cl}', \pi')$  in  $\mathbb{R}^d$  satisfying  $\Lambda_{\mathcal{S}'}^* = \Lambda_{\mathcal{S}}$ . Clearly, this reduction can be done in polynomial time. Thus, the FBCSD problem is  $\#P$ -hard for  $d \geq 3$ . In fact, with some efforts, one can make this result stronger by considering the FBCSD problem with respect to a balanced color pattern.

**Theorem 21** Let  $\mathcal{P}' = (\Delta'_1, \Delta'_2, \dots)$  be a balanced color pattern. Then for any fixed  $d$ , there exists a balanced color pattern  $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$  such that the CSD problem in  $\mathbb{R}^d$  with respect to  $\mathcal{P}$  is polynomial-time reducible to the FBCSD problem in  $\mathbb{R}^d$  with respect to  $\mathcal{P}'$ . In particular, the FBCSD problem in  $\mathbb{R}^d$  with respect to  $\mathcal{P}'$  is  $\#P$ -hard for  $d \geq 3$ .

### 3.2 Reduction to the CSD problem for $d = 2$

In this section, we study the FBCSD problem for  $d = 2$  and show that an instance of the FBCSD problem in  $\mathbb{R}^2$  can be reduced to  $O(n^2)$  instances of the CSD problem in  $\mathbb{R}^2$ . By combining this reduction with our algorithm given in Section 2.1, we directly obtain an  $O(n^4 \log^2 n)$ -time algorithm for the FBCSD problem in  $\mathbb{R}^2$ . For simplicity of exposition, we assume that  $S$  is in general position in  $\mathbb{R}^2$ , i.e., no three points are collinear.

We try to compute  $\Gamma_S^*$ . When computing  $\Gamma_S^*$ , we need to consider the realizations of  $S$  which contain no inter-color dominances with respect to some orthogonal basis of  $\mathbb{R}^2$  (these realizations are said to be *good*). We first establish a criterion for testing whether a realization is good. Recall that for a nonzero point  $x \in \mathbb{R}^d$ , the notation  $\bar{x}$  denotes the image of  $x$  in  $\mathbb{P}^{d-1}$  under the quotient map  $\mathbb{R}^d \setminus \{\mathbf{0}\} \rightarrow \mathbb{P}^{d-1}$ . For a subset  $A \subseteq S$ , we define  $L_A = \{a_i - a_j : a_i, a_j \in A \text{ and } \text{cl}(a_i) \neq \text{cl}(a_j)\} \subset \mathbb{P}^1$ . For two points  $l, l' \in \mathbb{P}^1$ , we denote by  $\theta(l, l')$  the angle between  $l$  and  $l'$  whose counterclockwise boundary is  $l$  and clockwise boundary is  $l'$  (when talking about angle we regard  $l$  and  $l'$  as lines in  $\mathbb{R}^2$  through the origin). Then we have the following observation.

**Lemma 22** A realization  $R$  of  $S$  is good iff  $L_R = \emptyset$  or there exists a unique  $l \in L_R$  such that  $\theta(l, l') > \frac{\pi}{2}$  for any  $l' \in L_R$  not equal to  $l$ .

Note that  $L_R = \emptyset$  iff  $R$  is monochromatic. Based on the above lemma, we now define a notion called *witness pair* as follows. Let  $R$  be a good but not monochromatic realization of  $S$ . Then by Lemma 22, there exists a unique  $l \in L_R$  such that  $\theta(l, l') > \frac{\pi}{2}$  for any  $l' \in L_R$  not equal to  $l$ . According to the definition of  $L_R$ , we must have  $l = \overline{a_i - a_j}$  for some  $a_i, a_j \in R$  with  $\text{cl}(a_i) \neq \text{cl}(a_j)$ . Note that the choice of  $a_i, a_j$  is not necessarily unique (though  $l$  is unique). Let  $Y$  be the set of all pairs  $(a_i, a_j)$  with  $a_i, a_j \in R$  satisfying  $\text{cl}(a_i) \neq \text{cl}(a_j)$  and  $l = \overline{a_i - a_j}$ . We claim that there exists a unique pair  $(a_{i^*}, a_{j^*}) \in Y$  such that for any  $(a_i, a_j) \in Y$  we have  $j^* \geq j$ . The existence is obvious, so it suffices to show the uniqueness. Indeed, if  $(a_i, a_j)$  and  $(a_{i'}, a_j)$  are two pairs in  $Y$ , then the points  $a_i, a_{i'}, a_j$  must be collinear in  $\mathbb{R}^2$ . However, because of the general position assumption for  $S$  (and hence for  $R$ ), we must have  $i = i'$ . It follows that for any  $j \in \{1, \dots, n\}$  there is at most one pair  $(a_i, a_j) \in Y$ , which further implies the uniqueness of  $(a_{i^*}, a_{j^*})$ . We define the pair  $(a_{i^*}, a_{j^*})$  as the *witness pair* of  $R$ , denoted by  $\text{wit}(R)$ . See Figure 5 for an example. Now it is clear that

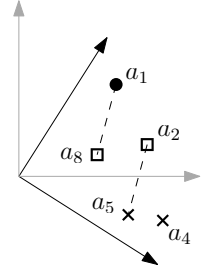


Figure 5: An example of witness pair.  $l = \overline{a_1 - a_8} = \overline{a_2 - a_5}$ .  $\text{wit}(R) = (a_1, a_8)$ .

$$\Gamma_S^* = Pr_{\text{mono}} + \sum_{i=1}^n \sum_{j=1}^n Pr_{i,j},$$

where  $Pr_{\text{mono}}$  is the probability that a realization  $R$  of  $S$  is monochromatic, and  $Pr_{i,j}$  is the probability that  $R$  is good (but not monochromatic) with  $\text{wit}(R) = (a_i, a_j)$ .

It is easy to compute  $Pr_{\text{mono}}$  in linear time. The problem remaining is how to compute  $Pr_{i,j}$  for all  $i, j \in \{1, \dots, n\}$ . Fix a pair  $(i^*, j^*)$ . Obviously, if  $\text{cl}(a_{i^*}) = \text{cl}(a_{j^*})$ , we immediately have  $Pr_{i^*, j^*} = 0$ . So suppose  $\text{cl}(a_{i^*}) \neq \text{cl}(a_{j^*})$ . We try to reduce the task of computing  $Pr_{i^*, j^*}$  to an instance of the CSD problem in  $\mathbb{R}^2$ . Let  $\mathbf{b}_1 = (a_{i^*} - a_{j^*}) / \|a_{i^*} - a_{j^*}\|_2$  be a unit vector of  $\mathbb{R}^2$ , and  $\mathbf{b}_2$  be another unit vector obtained by rotating  $\mathbf{b}_1$  clockwise with angle  $\frac{\pi}{2}$ . Clearly,  $B = (\mathbf{b}_1, \mathbf{b}_2)$  is an orthogonal basis of  $\mathbb{R}^2$ . We define  $n$  points  $a'_1, \dots, a'_n \in \mathbb{R}^2$  as follows. Let  $\delta$  be a small enough real number such that for any  $i, j \in \{1, \dots, n\}$  we have  $|\langle \mathbf{b}_2, a_i \rangle - \langle \mathbf{b}_2, a_j \rangle| > \delta$  unless  $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$ . Consider a specific index  $p \in \{1, \dots, n\}$ . If  $p \leq j^*$  and there exists  $q \leq j^*$  satisfying  $\text{cl}(a_p) \neq \text{cl}(a_q)$ ,  $a_p >_B a_q$ ,  $\langle \mathbf{b}_2, a_p \rangle = \langle \mathbf{b}_2, a_q \rangle$ , then we set the coordinates of  $a'_p$  in  $\mathbb{R}^2$  to be  $(\langle \mathbf{b}_2, a_p \rangle - \delta, \langle \mathbf{b}_1, a_p \rangle)$ . Otherwise, we set the coordinates of  $a'_p$  to be  $(\langle \mathbf{b}_2, a_p \rangle, \langle \mathbf{b}_1, a_p \rangle)$ . Based

on this, we can construct a colored stochastic dataset  $S' = (S', \text{cl}', \pi')$  in  $\mathbb{R}^2$  by defining  $S' = \{a'_1, \dots, a'_n\}$ ,  $\text{cl}'(a'_i) = \text{cl}(a_i)$  for all  $i \in \{1, \dots, n\}$ , and  $\pi'(a'_{i*}) = \pi'(a'_{j*}) = 1$ ,  $\pi'(a'_i) = \pi(a_i)$  for all  $i \in \{1, \dots, n\} \setminus \{i^*, j^*\}$ . We observe the following equation, which allows us to compute  $Pr_{i^*, j^*}$  by solving the instance  $\langle S' \rangle$  of the CSD problem in  $\mathbb{R}^2$ .

**Lemma 23**  $Pr_{i^*, j^*} = \pi(a_{i^*}) \cdot \pi(a_{j^*}) \cdot I_{S'}$ .

In this way, an instance of the FBCSD problem in  $\mathbb{R}^2$  is reduced to  $O(n^2)$  instances of the CSD problem in  $\mathbb{R}^2$ . By plugging in our  $O(n^2 \log^2 n)$  algorithm for solving the CSD problem in  $\mathbb{R}^2$ , we have the following result.

**Theorem 24** *The FBCSD problem in  $\mathbb{R}^2$  can be solved in  $O(n^4 \log^2 n)$  time.*

## References

- [1] P. Afshani, P.K. Agarwal, L. Arge, K.G. Larsen, and J.M. Phillips. (approximate) uncertain skylines. In *Proc. of the 14th ICDT*, pages 186–196. ACM, 2011.
- [2] P.K. Agarwal, B. Aronov, S. Har-Peled, J.M. Phillips, K. Yi, and W. Zhang. Nearest neighbor searching under uncertainty II. In *Proc. of the 32nd PODS*, pages 115–126. ACM, 2013.
- [3] P.K. Agarwal, S. Har-Peled, S. Suri, H. Yildiz, and W. Zhang. Convex hulls under uncertainty. In *Algorithms-ESA*, pages 37–48. Springer, 2014.
- [4] P.K. Agarwal, N. Kumar, S. Sintos, and S. Suri. Range-max queries on uncertain data. In *Proc. of the 35th SIGMOD/PODS*, pages 465–476. ACM, 2016.
- [5] M. Fink, J. Hershberger, N. Kumar, and S. Suri. Hyperplane separability and convexity of probabilistic point sets. In *Proc. of the 32nd SoCG*. ACM, 2016.
- [6] H.N. Gabow, J.L. Bentley, and R.E. Tarjan. Scaling and related techniques for geometry problems. In *Proc. of the 16th STOC*, pages 135–143. ACM, 1984.
- [7] P. Kamousi, T.M. Chan, and S. Suri. Stochastic minimum spanning trees in euclidean spaces. In *Proc. of the 27th SoCG*, pages 65–74. ACM, 2011.
- [8] P. Kamousi, T.M. Chan, and S. Suri. Closest pair and the post office problem for stochastic points. *Computational Geometry*, 47(2):214–223, 2014.
- [9] H-T. Kung, F. Luccio, and F.P. Preparata. On finding the maxima of a set of vectors. *Journal of the ACM*, 22(4):469–476, 1975.
- [10] C.H. Papadimitriou and M. Yannakakis. Multiobjective query optimization. In *Proc. of the 20th SIGMOD/PODS*, pages 52–59. ACM, 2001.
- [11] J. Pei, B. Jiang, X. Lin, and Y. Yuan. Probabilistic skylines on uncertain data. In *Proc. of the 33rd Intl. Conf. on VLDB*, pages 15–26. VLDB Endowment, 2007.
- [12] N. Robertson, D.P. Sanders, P. Seymour, and R. Thomas. Efficiently four-coloring planar graphs. In *Proc. of the 28th STOC*, pages 571–575. ACM, 1996.
- [13] S. Suri and K. Verbeek. On the most likely Voronoi Diagram and nearest neighbor searching. In *ISAAC*, pages 338–350. Springer, 2014.
- [14] S. Suri, K. Verbeek, and H. Yildiz. On the most likely convex hull of uncertain points. In *Algorithms-ESA*, pages 791–802. Springer, 2013.

- [15] W.T. Trotter. *Combinatorics and partially ordered sets: Dimension theory*, volume 6. JHU Press, 2001.
- [16] L.G. Valiant. Universality considerations in vlsi circuits. *IEEE Transactions on Computers*, 100(2):135–140, 1981.
- [17] M. Xia and W. Zhao. #3-regular bipartite planar vertex cover is #P-complete. In *Intl. Conf. on TAMC*, pages 356–364. Springer, 2006.
- [18] J. Xue, Y. Li, and R. Janardan. On the separability of stochastic geometric objects, with applications. In *Proc. of the 32nd SoCG*. ACM, 2016.
- [19] W. Zhang, X. Lin, Y. Zhang, M.A. Cheema, and Q. Zhang. Stochastic skylines. *ACM TODS*, 37(2):14, 2012.

# Appendix

## A Missing proofs

### A.1 Proof of Lemma 1

To see the “if” part, assume that  $Z(R)$  contains no inter-color dominances and  $y(a) > y(b)$  for any  $a \in Z(R)$ ,  $b \in R \setminus Z(R)$ . In this case, any two points in  $Z(R)$  cannot form an inter-color dominance. Also, any two points in  $R \setminus Z(R)$  cannot form an inter-color dominance for  $R \setminus Z(R)$  is monochromatic. It suffices to show that any  $a \in Z(R)$  and  $b \in R \setminus Z(R)$  cannot form an inter-color dominance. By assumption, we have  $y(a) > y(b)$ . But by the definition of  $Z(S)$ , we also have  $x(a) < x(b)$ . Thus,  $a$  and  $b$  do not dominate each other. To see the “only if” part, assume  $R$  contains no inter-color dominances. Since  $Z(R)$  is a subset of  $R$ , it also contains no inter-color dominances. Let  $a \in Z(R)$  and  $b \in R \setminus Z(R)$  be two points. As argued before, we have  $x(a) < x(b)$ . If  $\text{cl}(a) \neq \text{cl}(b)$ , then it is clear that  $y(a) > y(b)$  (otherwise  $(a, b)$  forms an inter-color dominance). The only remaining case is  $\text{cl}(a) = \text{cl}(b)$ . Since  $a \in Z(R)$ , by the definition of  $Z(R)$ , we may find a point  $o \in Z(R)$  such that  $x(a) < x(o) < x(b)$  and  $\text{cl}(o) \neq \text{cl}(a) = \text{cl}(b)$ . If  $y(a) < y(b)$ , then either  $y(a) < y(o)$  or  $y(o) < y(b)$ , i.e., either  $(a, o)$  or  $(o, b)$  forms an inter-color dominance. Because  $R$  contains no inter-color dominances, we must have  $y(a) > y(b)$ .

### A.2 Proof of Lemma 6

Suppose  $|V| = \{v_1, \dots, v_n\}$ . We construct the colored stochastic dataset  $\mathcal{S} = (S, \text{cl}, \pi)$  as follows. Define  $S = \{a_1, \dots, a_n\}$  where  $a_i = f(v_i) \in \mathbb{R}^d$  and set  $\text{cl}(a_i) = i$  (so  $\text{cl}$  is injective). Let  $S_1, \dots, S_c$  be the (disjoint) subsets of  $S$  corresponding to  $V_1, \dots, V_c$  respectively, i.e.,  $S_i = \{a_j : v_j \in V_i\}$ . Without loss of generality, we may assume  $S_1, \dots, S_c$  are all nonempty. For all points  $a \in S_i$ , we define  $\pi(a) = 4^{-n^{c-i+1}}$  (note that this real number can be represented in polynomial length). Then for all points  $a \in S \setminus (\bigcup_{i=1}^c S_i)$ , we define  $\pi(a) = \frac{1}{2}$ . With  $\mathcal{S}$  constructed above, we already have  $G_{\mathcal{S}} \cong G$ , since  $f$  is a DPE and all the points in  $S$  have distinct colors. It suffices to show how to “recover”  $\text{Ind}_{\Phi}(G)$  from  $\Gamma_{\mathcal{S}}$ . Equivalently, we have to compute, for every  $c$ -tuple  $\phi = (n_1, \dots, n_c)$  of integers where  $0 \leq n_i \leq |S_i|$ , the number of the subsets  $A \subseteq S$  containing no inter-color dominances and satisfying  $|A \cap S_i| = n_i$  for all  $i \in \{1, \dots, c\}$  (we use  $\mathcal{A}_{\phi}$  to denote the collection of these subsets). For each  $c$ -tuple  $\phi = (n_1, \dots, n_c)$  with  $0 \leq n_i \leq |S_i|$ , we notice that any  $A \in \mathcal{A}_{\phi}$  occurs as a realization of  $\mathcal{S}$  with probability

$$P_{\phi} = \frac{1}{2^{n-m}} \prod_{i=1}^c \left( \frac{1}{4^{n^{c-i+1}}} \right)^{n_i} \left( 1 - \frac{1}{4^{n^{c-i+1}}} \right)^{|S_i| - n_i},$$

where  $m = \sum_{i=1}^c |S_i|$ . Set  $N = \prod_{i=1}^c (|S_i| + 1)$ , then we have in total  $N$   $c$ -tuples  $\phi_1, \dots, \phi_N$  (of integers) to be considered ( $N$  is polynomial in  $n$  as  $c$  is constant). Suppose  $\phi_1, \dots, \phi_N$  are already sorted in lexicographical order from small to large. Our first key observation is that  $P_{\phi_i} > 2^n P_{\phi_{i+1}}$  for all  $i \in \{1, \dots, N-1\}$ . To see this, assume  $\phi_i = (n_1, \dots, n_c)$  and  $\phi_{i+1} = (n'_1, \dots, n'_c)$ . Note that  $\phi_1, \dots, \phi_N$  are sorted in lexicographical order, so there exists  $k \in \{1, \dots, c\}$  such that  $n_j = n'_j$  for all  $j < k$  and  $n'_k = n_k + 1$ . Then it is easy to see that

$$\frac{P_{\phi_i}}{P_{\phi_{i+1}}} \geq \frac{1 - 4^{-n^{c-k+1}}}{4^{-n^{c-k+1}}} \prod_{j=k+1}^c (4^{-n^{c-j+1}})^{|S_j|}.$$

If  $k = c$ , we already have  $P_{\phi_i} > 2^n P_{\phi_{i+1}}$ . For the case of  $k < c$ , since  $\sum_{j=k+1}^c |S_j| \leq n - 1$ , the above inequality implies that

$$\frac{P_{\phi_i}}{P_{\phi_{i+1}}} \geq \frac{(1 - 4^{-n^{c-k+1}}) \cdot 4^{-(n-1) \cdot n^{c-k}}}{4^{-n^{c-k+1}}} > 2^n.$$

With this observation in hand, we now consider how to compute  $|\mathcal{A}_{\phi_i}|$  for all  $i \in \{1, \dots, N\}$  from  $\Gamma_S$ . It is clear that

$$\Gamma_S = \sum_{i=1}^N P_{\phi_i} \cdot |\mathcal{A}_{\phi_i}|.$$

For  $j \in \{1, \dots, N\}$ , we set  $\gamma_j = \sum_{i=j+1}^N P_{\phi_i} \cdot |\mathcal{A}_{\phi_i}|$ . By the facts that  $P_{\phi_i} > 2^n P_{\phi_{i+1}}$  and  $\sum_{i=1}^N |\mathcal{A}_{\phi_i}| \leq 2^n$ , we can deduce  $P_{\phi_i} > \gamma_i$  for all  $i \in \{1, \dots, N\}$ . Then we are ready to compute  $|\mathcal{A}_{\phi_1}|, \dots, |\mathcal{A}_{\phi_N}|$  in order. Since  $P_{\phi_1} > \gamma_1$ ,  $|\mathcal{A}_{\phi_1}|$  must be the greatest integer that is smaller than or equal to  $\Gamma_S / P_{\phi_1}$ , and hence can be immediately computed. Suppose now  $|\mathcal{A}_{\phi_1}|, \dots, |\mathcal{A}_{\phi_{m-1}}|$  are already computed, and we consider  $|\mathcal{A}_{\phi_m}|$ . Via  $|\mathcal{A}_{\phi_1}|, \dots, |\mathcal{A}_{\phi_{m-1}}|$  and  $\Gamma_S$ , we may compute  $\gamma_{m-1}$ . Because  $P_{\phi_m} \geq \gamma_m$ ,  $|\mathcal{A}_{\phi_m}|$  must be the greatest integer that is smaller than or equal to  $\gamma_{m-1} / P_{\phi_m}$ , and hence can be computed directly. In this way, we are able to compute all  $|\mathcal{A}_{\phi_1}|, \dots, |\mathcal{A}_{\phi_N}|$  and equivalently  $\text{Ind}_{\Phi}(G)$  (in polynomial time). The statements in the lemma follow readily.

### A.3 Proof of Lemma 7

Since  $\mathcal{P}$  is balanced, we can find an constants  $c > 0$  such that  $n - \max \Delta_n \geq n^c$  for any sufficiently large  $n$ . Suppose  $G_S = (V \cup V', E)$  where  $|V| = n$  and  $|V'| = n'$ . We may write  $S = \{a_1, \dots, a_{n+n'}\}$  where  $a_1, \dots, a_n$  correspond to the vertices in  $V$  and  $a_{n+1}, \dots, a_{n+n'}$  correspond to those in  $V'$ . Because  $\text{cl}$  is injective (i.e., the points in  $S$  are of distinct colors), we have that  $a_1, \dots, a_n$  do not dominate each other, and the same holds for  $a_{n+1}, \dots, a_{n+n'}$ . Set  $N = \max\{2n + n', (n')^{1/c}\}$ . Now we construct  $S' = (S', \text{cl}', \pi')$  as follows. First, we pick a set  $A$  of  $N - (n + n')$  points in  $\mathbb{R}^d$  which do not dominate each other and do not form dominances with any points in  $S$ . Set  $S' = S \cup A$ , so  $S \subseteq S'$  and  $|S'| = N$ . The points in  $A$  are used as dummy points, and can never influences  $\Gamma_{S'}$  (since they are not involved in any dominances). With a little bit abuse of notation, we also use  $a_1, \dots, a_{n+n'}$  to denote the non-dummy points in  $S'$ . We then define  $\pi'$  as  $\pi'(a) = \pi(a)$  for  $a \in S$  and  $\pi'(a) = \frac{1}{2}$  for  $a \in A$ . It suffices to assign colors to the points in  $S'$ , i.e., define the coloring function  $\text{cl}'$ . Since we want  $\langle S' \rangle$  to be an instance of the CSD problem with respect to  $\mathcal{P}$ , the coloring  $\text{cl}'$  must induce the partition  $\Delta_N$  of  $N$ . Suppose  $\Delta_N = \{r_1, \dots, r_k\}$  (as a multi-set) where  $r_1 \geq \dots \geq r_k$ . Let  $l$  be the smallest integer such that  $\sum_{i=1}^l r_i \geq n$ . It is easy to see that  $\sum_{i=l+1}^k r_i \geq n'$ . Indeed, if  $l = 1$ , then we have

$$\sum_{i=2}^m r_i = N - \max \Delta_N \geq N^c \geq n'$$

by assumption. In the case of  $l > 1$ , we have that  $\sum_{i=1}^l r_i < 2n$  and thus  $\sum_{i=l+1}^k r_i > N - 2n \geq n'$ . This fact implies that we are able to define the coloring function  $\text{cl}'$  with image  $\{1, \dots, k\}$  such that (1) there are exactly  $r_i$  points in  $S'$  mapped to the color  $i$  by  $\text{cl}'$ , (2)  $\text{cl}'(a) \in \{1, \dots, l\}$  for any  $a \in \{a_1, \dots, a_n\}$ , (3)  $\text{cl}'(a) \in \{l+1, \dots, m\}$  for any  $a \in \{a_{n+1}, \dots, a_{n+n'}\}$ . With this  $\text{cl}'$ , we have that  $\text{cl}'(a_i) \neq \text{cl}'(a_j)$  for any  $i \in \{1, \dots, n\}$  and  $j \in \{n+1, \dots, n+n'\}$ . Therefore, if two points  $a_i, a_j \in S$  form an inter-color dominance in with respect to  $\text{cl}$ , then they also form an inter-color dominance with respect to  $\text{cl}'$ , and vice versa. Since the dummy points in  $A$  can never contribute inter-color dominances, we have  $\Gamma_{S'} = \Gamma_S$ , which completes the proof.

### A.4 Proof of Lemma 8

Fixing  $p, p' \in \{0, \dots, n\}$ , we denote by  $\mathcal{I}$  the collection of the independent sets  $I$  of  $G$  such that  $|I \cap V| = p$ ,  $|I \cap V'| = p'$ . Also, we denote by  $\mathcal{I}^*$  the collection of the independent sets  $I^*$  of  $G^*$  such that  $|I^* \cap V| = p$ ,  $|I^* \cap V'| = p'$ ,  $|I^* \cap U| = 3\lambda p$ ,  $|I^* \cap U'| = 3\lambda n - 3\lambda p$ . It suffices to establish an one-to-one correspondence between  $\mathcal{I}$  and  $\mathcal{I}^*$ . Let  $I \in \mathcal{I}$  be an element. If  $e = (v, v') \in E$  is an edge of  $G$  (where  $v \in V$  and  $v' \in V'$ ), we say  $e$  is of Type-1 if  $v \in I$  (and hence  $v' \notin I$ ), otherwise of Type-2. Recall that for each  $e \in E$ ,  $U_e$  (resp.,  $U'_e$ ) denotes the set of the  $\lambda$  vertices in  $U$  (resp.,  $U'$ ) which are inserted to the edge  $e$ . Now let  $I^*$  be the set consists of the vertices in  $I$ , the vertices in  $U_e$  for all Type-1 edges  $e$ , and the vertices in  $U'_e$  for all Type-2 edges  $e$ . Clearly,  $I^*$  is an independent set of  $G^*$ . Furthermore, by the definition of  $\mathcal{I}$  and the fact that  $G$



is 3-regular, we know that  $G$  has  $3p$  Type-1 edges and  $3n - 3p$  Type-2 edges. It follows that  $|I^* \cap V| = p$ ,  $|I^* \cap V'| = p'$ ,  $|I^* \cap U| = 3\lambda p$ ,  $|I^* \cap U'| = 3\lambda n - 3\lambda p$ . Thus,  $I^* \in \mathcal{I}^*$ . By mapping  $I$  to  $I^*$ , we obtain a map from  $\mathcal{I}$  to  $\mathcal{I}^*$ , which is obviously injective. To see it is surjective, let  $I^* \in \mathcal{I}^*$  be an element. Set  $I = I^* \cap (V \cup V')$ . We claim that  $I \in \mathcal{I}$  and  $I$  is mapped to  $I^*$  by our map defined above. First, since  $I^*$  is an independent set of  $G^*$ , we must have  $|I^* \cap (U_e \cup U'_e)| \leq \lambda$  for any edge  $e = (v, v') \in E$  of  $G$  (with equality only if at least one of  $v$  and  $v'$  is in  $I$ ). But  $|I^* \cap (U \cup U')| = 3\lambda n = \lambda|E|$ , which implies  $|I^* \cap (U_e \cup U'_e)| = \lambda$  for all  $e \in E$ . It follows that for every edge  $e = (v, v') \in E$ ,  $v$  and  $v'$  are not included in  $I$  simultaneously, i.e.,  $I$  is an independent set of  $G$ . In addition,  $|I \cap V| = |I^* \cap V| = p$ ,  $|I \cap V'| = |I^* \cap V'| = p'$ . Therefore,  $I \in \mathcal{I}$ . To see  $I$  is mapped to  $I^*$ , we apply again the fact that  $|I^* \cap (U_e \cup U'_e)| = \lambda$  for any  $e \in E$ . Based on this, we further observe that for any  $e \in E$ , either  $U_e \subseteq I^*$  or  $U'_e \subseteq I^*$  (since  $I^*$  is an independent set of  $G^*$ ). As before, we say an edge  $e = (v, v') \in E$  (with  $v \in V$  and  $v' \in V'$ ) is of Type-1 if  $v \in I$ , otherwise of Type-2. Note that if an edge  $e \in E$  is of Type-1, we must have  $U_e \subseteq I^*$  (and then  $I^* \cap U'_e = \emptyset$ ). Since  $G$  has  $3p$  Type-1 edges,  $|I^* \cap U| \geq 3\lambda p$ . But in fact  $|I^* \cap U| = 3\lambda p$  as  $I^* \in \mathcal{I}^*$ . So the only possibility is that  $U_e \subseteq I^*$  (and  $I^* \cap U'_e = \emptyset$ ) for all Type-1 edges  $e$  and  $U'_e \subseteq I^*$  (and  $I^* \cap U_e = \emptyset$ ) for all Type-2 edges  $e$ . As a result,  $I$  is mapped to  $I^*$  and  $|\mathcal{I}| = |\mathcal{I}^*|$ , completing the proof.

## A.5 Proof of Lemma 10

The “if” part is obvious, because  $\underline{\max}(A)$  clearly dominates every point in  $A$  for any (finite)  $A \subset \mathbb{R}^d$  with  $|A| \geq 2$  (note that  $|A_{w'}| \geq 2$  for any  $w' \in V' \cup U'$ ). It suffices to prove the “only if” part. For a point  $p \in \mathbb{R}^3$ , we denote by  $H_p$  the set of the points on the plane  $H$  which are dominated by  $p$ . We first observe that if  $H_p \neq \emptyset$ , then the preimage  $\psi^{-1}(H_p)$  of  $H_p$  under  $\psi$  (which is a region in  $\mathbb{R}^2$ ) must be a (closed) right-angled isosceles triangle in  $\mathbb{R}^2$  whose hypotenuse is horizontal (we call this kind of triangles *standard triangles*). To see this, assume  $p = (x_p, y_p, z_p)$  and  $H_p \neq \emptyset$  (this is equivalent to saying  $x_p + y_p + z_p > 0$ ). Then  $\psi^{-1}(H_p)$  consists of all the points  $(x, y) \in \mathbb{R}^2$  satisfying  $x + y \leq x_p$ ,  $y - x \leq y_p$ ,  $y \geq -z_p/2$ , and hence is a standard triangle. Furthermore, it is easy to see that if  $p = \underline{\max}(A)$  for a finite set  $A \subset H$  with  $|A| \geq 2$ , then  $H_p \neq \emptyset$  and  $\psi^{-1}(H_p)$  is the minimal standard triangle containing  $\psi^{-1}(A)$  (by “minimal” we mean that any standard triangle containing  $\psi^{-1}(A)$  is a superset of  $\psi^{-1}(H_p)$ , both as subsets of  $\mathbb{R}^2$ , see Figure 6). Therefore, we only need to show that for any vertex  $w' \in V' \cup U'$ , the minimal standard triangle containing  $\psi^{-1}(A_{w'}) = \varphi(\text{Adj}_{w'})$  does not contain  $\varphi(w)$  for any vertex  $w \in (V \cup U) \setminus \text{Adj}_{w'}$ . We consider two cases,  $w' \in V'$  and  $w' \in U'$ .

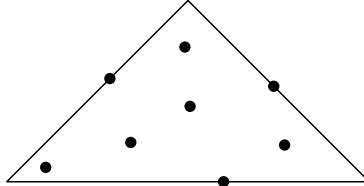


Figure 6: The minimal standard triangle in  $\mathbb{R}^2$  containing a set of points.

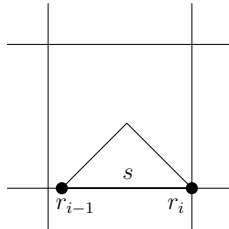


Figure 7: The case that  $s$  is horizontal.

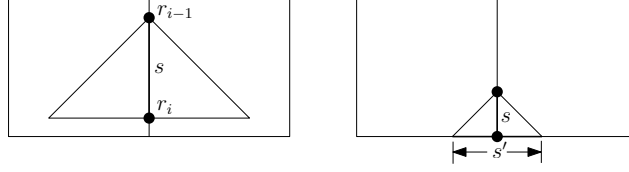


Figure 8: The case that  $s$  is vertical.

In the case of  $w' \in V'$ ,  $\text{Adj}_{w'}$  consists of three vertices (for  $G$  is 3-regular) in  $U$ , say  $w_1, w_2, w_3$ . Recall that  $g$  is the OGD of  $G$  used in constructing the map  $\varphi$ . By retrospectively constructing our construction of  $\varphi$ , we see that each of  $\varphi(w_1), \varphi(w_2), \varphi(w_3)$  has distance 0.01 from  $g(w')$ . On the other hand, one can easily verify that for any vertex  $w \in (V \cup U) \setminus \text{Adj}_{w'}$ ,  $\varphi(w)$  is “far away” from  $g(w')$  (more precisely, with distance at least 0.3). Therefore, the minimal standard triangle containing  $\varphi(w_1), \varphi(w_2), \varphi(w_3)$  does not contain  $\varphi(w)$  for any vertex  $w \in (V \cup U) \setminus \text{Adj}_{w'}$ .

In the case of  $w' \in U'$ , we may assume  $w' \in U'_e$  for some edge  $e = (v, v') \in E$  of  $G$ . Then  $\text{Adj}_{w'}$  consists of two vertices in  $\{v\} \cup U_e$ , say  $w_1, w_2$ . Recall that  $P_e$  is the set of the  $\lambda$  points chosen on the curve  $g(e)$  for sake of defining  $\varphi(u)$  for  $u \in U_e$ . As before, we suppose  $P_e = \{r_1, \dots, r_\lambda\}$  where  $r_1, \dots, r_\lambda$  are sorted in the order they appear on the curve  $g(e)$  (from  $g(v)$  to  $g(v')$ ). For convenience, set  $r_0 = g(v)$ . Then we may assume  $\varphi(w_1) = r_{i-1}$  and  $\varphi(w_2) = r_i$  for some  $i \in \{1, \dots, \lambda\}$ . Let  $s = \overline{r_{i-1}r_i}$  be the segment in  $\mathbb{R}^2$  with endpoints  $r_{i-1}$  and  $r_i$ , and  $\Delta$  be the minimal standard triangle containing  $r_{i-1}$  and  $r_i$ . Since all the grid points on  $g(e)$  are included in  $P_e$ ,  $s$  must be a horizontal or vertical segment contained in  $g(e)$ . Furthermore, the interior of  $s$  does not contain  $\varphi(w)$  for any vertex  $w \in V \cup U$  and in particular does not contain any grid points. We discuss two cases separately:  $s$  is horizontal and  $s$  is vertical. Recall that  $K = (\mathbb{Z} \times \mathbb{R}) \cup (\mathbb{R} \times \mathbb{Z}) \subset \mathbb{R}^2$  is the grid. If  $s$  is horizontal, then  $\Delta$  is just the standard triangle having  $s$  as its hypotenuse (see Figure 7). In this case, we have  $\Delta \cap K = s$ , which implies that  $\Delta$  does not contain  $\varphi(w)$  for any vertex  $w \in (V \cup U) \setminus \{w_1, w_2\}$ . For the case that  $s$  is vertical, assume that  $r_{i-1}$  is the top endpoint and  $r_i$  is the bottom one. Then  $r_{i-1}$  is the right-angled vertex of  $\Delta$ , and  $r_i$  is the midpoint of the hypotenuse of  $\Delta$ . If  $r_i$  is not a grid point, we again have  $\Delta \cap K = s$  and thus we are done (see the left part of Figure 8). If  $r_i$  is a grid point, the distance between  $r_{i-1}$  and  $r_i$  must be 0.3, by our construction of  $P_e$ . In this situation,  $\Delta \cap K$  consists of  $s$  and a horizontal segment  $s'$  of length 0.6 which is the hypotenuse of  $\Delta$  (see the right part of Figure 8). We claim that  $\varphi(w)$  is not on  $s'$  for any vertex  $w \in (V \cup U) \setminus \{w_2\}$ . Indeed, by our construction of  $\varphi$ , if  $\varphi(w)$  is in the interior of some unit horizontal segment, then  $\varphi(w)$  is either with distance 0.01 from  $g(v')$  for some  $v' \in V'$  or with distance at least 0.4 from any grid point. In each of the cases,  $\varphi(w)$  is “far away” from  $r_i$  (more precisely, with distance at least 0.4). But any point on  $s'$  has distance at most 0.3 from  $r_i$ . Therefore,  $\varphi(w)$  is not on  $s'$ . It immediately follows that  $\Delta$  does not contain  $\varphi(w)$  for any vertex  $w \in (V \cup U) \setminus \{w_1, w_2\}$ , which completes the proof.

## A.6 Proof of Theorem 12

Suppose  $n = |V \cup V'|$ . Let  $h : V \rightarrow \{1, \dots, k\}$  be a semi-discrete  $k$ -halfcoloring of  $G$  (on  $V$ ). We show  $\dim(G) \leq 2k$  by explicitly constructing a DPE  $f : V \cup V' \rightarrow \mathbb{R}^{2k}$  of  $G$ . For  $i \in \{1, \dots, k\}$ , we define  $V_i = h^{-1}(\{i\}) \subseteq V$  (i.e.,  $V_i$  consists of the vertices in  $V$  colored with color  $i$  by  $h$ ) and define  $G_i$  as the subgraph of  $G$  with the vertex set  $V_i \cup V'$ . We first construct  $k$  functions  $f_1, \dots, f_k : V \cup V' \rightarrow \mathbb{R}^2$ , and then obtain the DPE  $f$  by identifying  $\mathbb{R}^{2k}$  with  $(\mathbb{R}^2)^k$  and “combining” the functions  $f_1, \dots, f_k$ , i.e., setting

$$f(v) = (f_1(v), \dots, f_k(v))$$

for all  $v \in V \cup V'$ . Fixing  $p \in \{1, \dots, k\}$ , we describe the construction of  $f_p$ . Suppose the graph  $G_p$  consists of  $m$  connected components. For each  $i \in \{1, \dots, m\}$ , let  $C_i$  be the set of the vertices in the  $i$ -th connected component of  $G_p$ . Also, for each  $i \in \{1, \dots, m\}$ , let

$$B_i = \{(x, y) \in \mathbb{R}^2 : i-1 < x < i, m-i < y < m-i+1\}$$

be an open box in  $\mathbb{R}^2$  (see the left part of Figure 9). The function  $f_p$  to be constructed maps the vertices in  $C_i$  to points in  $B_i$  as follows. Since  $h$  is semi-discrete, we know that  $|C_i \cap V| \leq 2$ . If  $|C_i \cap V| = 0$ , then  $C_i$  only contains an isolated vertex  $v' \in V'$ , and we set  $f_p(v')$  to be an arbitrary point in  $B_i$ . If  $|C_i \cap V| = 1$ , let  $v$  be the only vertex in  $C_i \cap V$  and suppose  $C_i \cap V' = \{v'_1, \dots, v'_r\}$ . In this case, we set  $f_p(v'_1), \dots, f_p(v'_r)$  to be a sequence of  $r$  points in  $B_i$  with increasing  $x$ -coordinates and decreasing  $y$ -coordinates, and  $f_p(v)$  to be an arbitrary point in  $B_i$  dominated by all of  $f_p(v'_1), \dots, f_p(v'_r)$ . See the middle part of Figure 9 for an intuitive illustration for this case. If  $|C_i \cap V| = 2$ , let  $v_1, v_2$  be the two vertices in  $C_i \cap V$  and again suppose  $C_i \cap V' = \{v'_1, \dots, v'_r\}$ . We may assume that the vertices in  $C_i \cap V'$  adjacent to  $v_1$  (resp.,  $v_2$ ) are exactly  $v'_1, \dots, v'_\alpha$  (resp.,  $v'_\beta, \dots, v'_r$ ) for some  $\alpha, \beta \in \{1, \dots, r\}$  with  $\alpha \geq \beta$  (if not, one can easily relabel the points to achieve this). Again, we set  $f_p(v'_1), \dots, f_p(v'_r)$  to be a sequence of  $r$  points in  $B_i$  with increasing  $x$ -coordinates and decreasing  $y$ -coordinates. Then we set  $f_p(v_1)$  to be a point in  $B_i$  which is dominated by exactly  $f_p(v'_1), \dots, f_p(v'_\alpha)$ , and set  $f_p(v_2)$  to be a point in  $B_i$  which is dominated by exactly  $f_p(v'_\beta), \dots, f_p(v'_r)$ . Note that we can definitely find such two points, since  $f_p(v'_1), \dots, f_p(v'_r)$  have increasing  $x$ -coordinates and decreasing  $y$ -coordinates. In addition, by carefully determining the locations of  $f_p(v_1)$  and  $f_p(v_2)$  in  $B_i$ , we may further require that  $f_p(v_1)$  and  $f_p(v_2)$  do not dominate each other. See the right part of Figure 9 for an intuitive illustration for this case. After considering all  $C_i$ , the function  $f_p$  is defined for all vertices in  $V_p \cup V'$  (which is the vertex set of  $G_p$ ). So it suffices to define  $f_p$  on  $V \setminus V_p$ . For each  $v \in V \setminus V_p$ , we simply set  $f_p(v)$  to be an arbitrary point in the box  $[-N, -N + 1] \times [-N, -N + 1]$  for a sufficiently large integer  $N > 10n$  (recall that  $n = |V \cup V'|$ ), which completes the construction of  $f_p$ . We observe that  $f_p$  has the following properties.

- (1) For any  $v \in V$  and  $w \in V_p$ ,  $f_p(v) \not\prec f_p(w)$ .
- (2) For any  $v' \in V'$ ,  $f_p(v')$  is not dominated by any point in the image of  $f_p$ .
- (3) For any  $v \in V_p$  and  $v' \in V'$ ,  $f_p(v') > f_p(v)$  iff  $v$  and  $v'$  are adjacent in  $G$ .

We do the same thing for all  $p \in \{1, \dots, k\}$  and obtain the functions  $f_1, \dots, f_k$ . As mentioned before, we then define  $f : V \cup V' \rightarrow \mathbb{R}^{2k}$  as  $f(v) = (f_1(v), \dots, f_k(v))$ . We now prove that  $f$  is a DPE of  $G$ . First, for any  $v \in V$ , we claim that  $f(v)$  does not dominate any point in the image of  $f$ . Indeed,  $f(v) \not\prec f(v')$  for any  $v' \in V'$ , since  $f_1(v')$  is not dominated by any point in the image of  $f_1$  by the property (2) above. Also,  $f(v) \not\prec f(w)$  for any  $w \in V$ , since  $f_p(v) \not\prec f_p(w)$  for  $p = h(w)$  by the property (1) above. Second, for any  $v' \in V'$ , we have that  $f(v')$  is not dominated by any point in the image of  $f$ , simply because  $f_1(v')$  is not dominated by any point in the image of  $f_1$  by the property (2) above. Finally, consider two vertices  $v \in V$  and  $v' \in V'$ . We claim that  $f(v') > f(v)$  iff  $v$  and  $v'$  are adjacent in  $G$ . If  $v$  and  $v'$  are adjacent, then  $f_i(v') > f_i(v)$  for all  $i \in \{1, \dots, k\}$  by the property (3) above, and hence  $f(v') > f(v)$ . If  $v$  and  $v'$  are not adjacent, then  $f_p(v') \not\prec f_p(v)$  for  $p = h(v)$  by the property (3) above, and hence  $f(v') \not\prec f(v)$ . In sum, we have  $f(v') > f(v)$  iff  $v \in V$ ,  $v' \in V'$ ,  $v$  and  $v'$  are adjacent in  $G$ . Therefore,  $f$  is a DPE of  $G$  to  $\mathbb{R}^{2k}$ . Clearly,  $f$  can be constructed in polynomial time if the  $k$ -halfcoloring  $h$  is provided, which completes the proof of the first part of the theorem.

Next, we prove the second part of the theorem. Again, let  $h : V \rightarrow \{1, \dots, k\}$  be a semi-discrete  $k$ -halfcoloring of  $G$  (on  $V$ ). Suppose  $\chi_h(v') < k$  for all  $v' \in V'$ . If  $k = 1$ , then  $\chi_h(v') = 0$  for all  $v' \in V'$ , which implies that  $G$  has no edges and thus the statement is trivial (any constant map  $f : V \cup V' \rightarrow \mathbb{R}$  is a DPE of  $G$ ). So assume  $k \geq 2$ . We show  $\dim(G) \leq 2k - 1$  by explicitly constructing a DPE  $f : V \cup V' \rightarrow \mathbb{R}^{2k-1}$  of

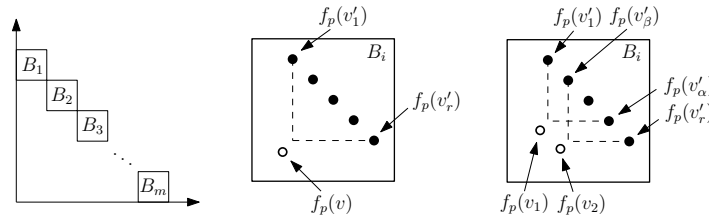


Figure 9: A local structure of  $f_p$  in the box  $B_i$ .

$G$ . In the same way as before, we define the functions  $f_1, \dots, f_k : V \cup V' \rightarrow \mathbb{R}^2$ . But we need a different way to define  $f$ . To this end, we first construct  $k-1$  functions  $f'_1, \dots, f'_{k-1} : V \cup V' \rightarrow \mathbb{R}^2$  based on  $f_1, \dots, f_k$  as follows. Fixing  $p \in \{1, \dots, k-1\}$ , we describe the construction of  $f'_p$ . For all  $v \in V \setminus V_k$ , we set  $f'_p(v) = f_p(v)$ . For all  $v \in V_k$ , we set  $f'_p(v) = f_k(v) - (n, n)$ , that is, if  $f_k(v) = (x, y) \in \mathbb{R}^2$  then  $f'_p(v) = (x - n, y - n)$ . Now consider the vertices in  $V'$ . If a vertex  $v' \in V'$  is “adjacent” to the color  $p$  (recall that  $v'$  is said to be “adjacent” to the color  $p$  if there exists  $v \in V$  adjacent to  $v'$  with  $h(v) = p$ ), then we set  $f'_p(v') = f_p(v')$ , otherwise  $f'_p(v') = f_k(v') - (n, n)$ . By doing this for all  $p \in \{1, \dots, k-1\}$ , we complete constructing  $f'_1, \dots, f'_{k-1}$ . However, if we “combine”  $f'_1, \dots, f'_{k-1}$ , we only obtain a map  $V \cup V' \rightarrow \mathbb{R}^{2k-2}$  which is not guaranteed to be a DPE. So the last ingredient needed for defining  $f$  is a function  $\rho : V \cup V' \rightarrow \mathbb{R}$ . The definition of  $\rho$  is quite simple. We set  $\rho(v) = 1$  for all  $v \in V \setminus V_k$ , and  $\rho(v) = 3$  for all  $v \in V_k$ . For  $v' \in V'$ , if  $v'$  is “adjacent” to the color  $k$  or  $\chi_h(v') = 0$ , then we set  $\rho(v') = 4$ , otherwise  $\rho(v') = 2$ . Finally,  $f : V \cup V' \rightarrow \mathbb{R}^{2k-1}$  is defined by identifying  $\mathbb{R}^{2k-1}$  with  $(\mathbb{R}^2)^{k-1} \times \mathbb{R}$  and “combining” the functions  $f'_1, \dots, f'_{k-1}, \rho$ , i.e., setting

$$f(v) = (f'_1(v), \dots, f'_{k-1}(v), \rho(v))$$

for all  $v \in V \cup V'$ . We need to verify that  $f$  is truly a DPE of  $G$  to  $\mathbb{R}^{2k-1}$ .

First, we show that for any  $v \in V$ ,  $f(v)$  does not dominate any point in the image of  $f$ . Let  $v \in V$  be a vertex. We consider two cases,  $v \in V \setminus V_k$  and  $v \in V_k$ . In the case of  $v \in V \setminus V_k$ , we first notice that  $f(v) \not\triangleright f(w)$  for any  $w \in V_k \cup V'$ , simply because  $\rho(v) < \rho(w)$ . To see this  $f(v) \not\triangleright f(w)$  for any  $w \in V \setminus V_k$ , set  $p = h(w) \neq k$ . Then  $f'_p(v) = f_p(v)$  does not dominate  $f'_p(w) = f_p(w)$  by the property (1) above, and hence  $f(v) \not\triangleright f(w)$ . In the case of  $v \in V_k$ , we first claim that  $f(v) \not\triangleright f(w)$  for any  $w \in V$ . If  $w \notin V_k$ , then by setting  $p = h(w) \neq k$  we have  $f'_p(v) = f_k(v) - (n, n)$  does not dominate  $f'_p(w) = f_p(w)$ , which implies  $f(v) \not\triangleright f(w)$ . If  $w \in V_k$ , then  $f'_1(v) = f_k(v) - (n, n)$  does not dominate  $f'_1(w) = f_k(w) - (n, n)$  since  $f_k(v) \not\triangleright f_k(w)$  by the property (1) above, which also implies  $f(v) \not\triangleright f(w)$ . It suffices to show that  $f(v) \not\triangleright f(v')$  for any  $v' \in V'$ . Indeed, we have either  $f'_1(v') = f_1(v')$  or  $f'_1(v') = f_k(v') - (n, n)$ . In each case,  $f'_1(v) = f_k(v) - (n, n)$  does not dominate  $f'_1(v')$  (the former case is obvious and the latter case follows from the property (2) above). Thus  $f(v) \not\triangleright f(v')$ .

Second, we show that for any  $v' \in V'$ ,  $f(v')$  is not dominated by any point in the image of  $f$ . Let  $v' \in V'$  be a vertex. By the argument above, it suffices to verify that  $f(w') \not\triangleright f(v')$  for any  $w' \in V'$ . If  $v'$  is “adjacent” to some color  $p \in \{1, \dots, k-1\}$ , then we are done because  $f'_p(v') = f_p(v')$  is not dominated by  $f'_p(w')$  for any  $w' \in V'$ . Suppose  $v'$  is not “adjacent” to any color in  $\{1, \dots, k-1\}$ . In this case, we must have  $\rho(v') = 4$  and  $f'_i(v') = f_k(v') - (n, n)$  for all  $i \in \{1, \dots, k-1\}$ . We first notice that  $f(w') \not\triangleright f(v')$  for any  $w' \in V'$  such that  $\chi_h(w') > 0$  and  $w'$  is not “adjacent” to the color  $k$ , simply because  $\rho(w') = 2 < \rho(v')$ . Then we consider the case that  $w' \in V'$  is “adjacent” to the color  $k$  or  $\chi_h(w') = 0$ . By the assumption  $\chi_h(w') < k$ , we know that  $w'$  cannot be “adjacent” to all the  $k$  colors. In other words, if  $w'$  is “adjacent” to the color  $k$  or  $\chi_h(w') = 0$ ,  $w'$  must miss some color in  $\{1, \dots, k-1\}$ . Without loss of generality, we may assume  $w'$  is not “adjacent” to the color 1. Thus,  $f'_1(v') = f_k(v') - (n, n)$  is not dominated by  $f'_1(w') = f_k(w') - (n, n)$  by the property (2) above, and hence  $f(w') \not\triangleright f(v')$ .

Finally, we show that for any  $v \in V$  and  $v' \in V'$ ,  $f(v') > f(v)$  iff  $v$  and  $v'$  are adjacent in  $G$ . Let  $v \in V$  and  $v' \in V'$  be two vertices. If  $v$  and  $v'$  are adjacent in  $G$ , one can easily verify (by checking various cases) that  $\rho(v') > \rho(v)$  and  $f'_i(v')$  dominates  $f'_i(v)$  for all  $i \in \{1, \dots, k-1\}$ , which implies  $f(v') > f(v)$ . Now suppose  $v$  and  $v'$  are not adjacent in  $G$ . We consider two cases,  $v \in V \setminus V_k$  and  $v \in V_k$ . In the case of  $v \in V \setminus V_k$ , set  $p = h(v) \neq k$ . Then  $f'_p(v) = f_p(v)$ . Besides, we have either  $f'_p(v') = f_p(v')$  or  $f'_p(v') = f_k(v') - (n, n)$ . For the former,  $f'_p(v') \not\triangleright f'_p(v)$  follows from the property (3) above, while for the latter  $f'_p(v') \not\triangleright f'_p(v)$  follows obviously. Thus,  $f(v') \not\triangleright f(v)$ . In the case of  $v \in V_k$ , we have  $f'_i(v) = f_k(v) - (n, n)$  for all  $i \in \{1, \dots, k\}$  and  $\rho(v) = 3$ . If  $v'$  is not “adjacent” to the color  $k$  and  $\chi_h(v') > 0$ , then  $\rho(v') = 2 < \rho(v)$  and hence  $f(v') \not\triangleright f(v)$ . If  $v'$  is “adjacent” to the color  $k$  or  $\chi_h(v') = 0$ , then as argued before  $v'$  must miss some color in  $\{1, \dots, k-1\}$ . Without loss of generality, we may assume  $w'$  is not “adjacent” to the color 1. Thus,  $f'_1(v') = f_k(v') - (n, n)$  does not dominate  $f'_1(v) = f_k(v) - (n, n)$  by the property (3) above, which implies  $f(v') \not\triangleright f(v)$ .

In sum, two vertices in  $G$  share a common edge iff their images under  $f$  form a dominance. Therefore,  $f$  is a DPE of  $G$  to  $\mathbb{R}^{2k-1}$ . It is clear that the construction of  $f$  can be done in polynomial time if the  $k$ -halfcoloring  $h$  is provided.

## A.7 Proof of Lemma 13

Let  $G = (V \cup V', E)$  be a 3-regular planar bipartite graph. As before, we define the graph  $G' = (V, E')$  by setting  $E' = \{(a, b) : a \sim b \text{ in } G\}$ . Then a discrete  $k$ -halfcoloring of  $G$  on  $V$  corresponds to a (conventional)  $k$ -coloring of  $G'$  satisfying that no two adjacent vertices share the same color. We first show that  $G'$  is planar. Fix a planar drawing  $\varphi$  of  $G$ . Let  $v' \in V'$  be a vertex. Since  $G$  is 3-regular,  $v'$  must be adjacent to three vertices  $v_1, v_2, v_3 \in V$ . We now delete  $v'$  as well as its three adjacent edges from  $G$  and add three new edges  $(v_1, v_2), (v_2, v_3), (v_3, v_1)$  to  $G$ . We claim that the resulting graph is still planar. Indeed, in the drawing  $\varphi$ , after we remove  $\varphi(v')$  and its adjacent edges,  $\varphi(v_1), \varphi(v_2), \varphi(v_3)$  will share a common face, which is the one previously containing  $\varphi(v')$ . So we can draw the edges  $(v_1, v_2), (v_2, v_3), (v_3, v_1)$  inside this face along with the image of the deleted edges (see Figure 10). In this way, we keep deleting the vertices in  $V'$  (as well as the adjacent edges) and adding new edges. In this process, the planarity of the graph always keeps. Until all the vertices in  $V'$  are deleted, the resulting graph, which is still planar, is nothing but  $G'$ , as two vertices  $u, v \in V$  are connected (in the resulting graph) iff  $u \sim v$  in  $G$ . By applying the well-known Four Color Theorem, we know that  $G'$  is 4-colorable. Furthermore, to find a 4-coloring for  $G'$  can be done in quadratic time using the approach in [12]. As a result, a discrete 4-halfcoloring of  $G$  can be computed in polynomial time, completing the proof.

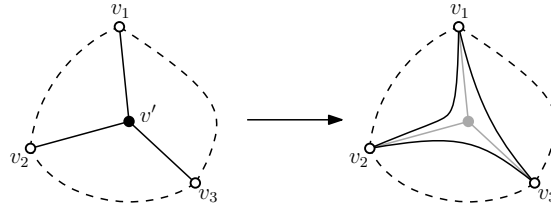


Figure 10: Deleting a vertex and adding three new edges.

## A.8 Proof of Theorem 14

We show that for any  $i, j \in \{1, \dots, n\}$  with  $i < j$ ,

$$\Pr[E_{i,j}] \cdot |Est_{i,j} - Cond_{i,j}| < \frac{\varepsilon}{n^2} \Lambda_S \quad (3)$$

with probability  $1 - O(e^{-n})$ . As long as this is true, by using union bound, we can immediately conclude that  $|\Lambda - \Lambda_S| < \varepsilon \Lambda_S$  with probability at least  $2/3$ , which completes the proof. Consider a realization  $R$  of  $\mathcal{S}_{i-1}$ . Clearly, the probability that  $R \cup \{a_i, a_j\}$  contains inter-color dominances is nothing but  $Cond_{i,j}$ . Therefore, by Hoeffding's inequality and the definition of  $Est_{i,j}$ , we have that

$$\Pr \left[ |Est_{i,j} - Cond_{i,j}| \geq \frac{\varepsilon}{n^2} \right] \leq 2e^{-2N\varepsilon^2/n^4} = 2e^{-2n}.$$

If  $\Pr[E_{i,j}] \leq \Lambda_S$ , we are done because the above already implies that Inequality 3 holds with probability  $1 - O(e^{-n})$ . So assume  $\Pr[E_{i,j}] > \Lambda_S$ . Note that  $\pi(a_i) \cdot \pi(a_j) \geq \Pr[E_{i,j}]$ , which implies  $\pi(a_i) \cdot \pi(a_j) > \Lambda_S$ . We claim that  $Cond_{i,j} = 0$ . It suffices to show that for any realization  $R$  of  $\mathcal{S}_{i-1}$ ,  $R \cup \{a_i, a_j\}$  contains no inter-color dominances. Let  $a_p, a_q \in R \cup \{a_i, a_j\}$  be two distinct points. Assume  $cl(a_p) \neq cl(a_q)$  and  $a_p > a_q$ . Then we must have  $\Lambda_S \geq \pi(a_p) \cdot \pi(a_q)$  because a realization of  $\mathcal{S}$  does contain inter-color dominances if it includes both  $a_p$  and  $a_q$ . However, recall that  $\pi(a_1) \geq \dots \geq \pi(a_n)$ . Thus,  $\pi(a_p) \cdot \pi(a_q) \geq \pi(a_i) \cdot \pi(a_j) > \Lambda_S$ , which gives us a contradiction. Since  $Cond_{i,j} = 0$ ,  $Est_{i,j}$  is for sure 0. It follows that Inequality 3 holds with probability 1 in this case. As a result,  $(1 - \varepsilon)\Lambda_S < \Lambda < (1 + \varepsilon)\Lambda_S$  with probability at least  $2/3$ .

## A.9 Proof of Lemma 15

Fixing  $p \in \{1, \dots, d\}$ , we determine the  $p$ -th coordinates of  $\hat{a}_1, \dots, \hat{a}_n$  as follows. For all  $i \in \{1, \dots, n\}$ , define a triple  $\phi_i = (\gamma_i, \sigma_i, i)$  where  $\gamma_i$  is the  $p$ -th coordinate of  $a_i$  and  $\sigma_i$  is the sum of the  $d$  coordinates of  $a_i$ . Then we sort all  $\phi_i$  in lexicographic order from small to large, and suppose  $\phi_{i_1}, \dots, \phi_{i_n}$  is the resulting sorted sequence. We have  $\phi_{i_1} < \dots < \phi_{i_n}$  under lexicographic order, since there exist no ties. Now we simply set the  $p$ -th coordinates of  $\hat{a}_{i_1}, \dots, \hat{a}_{i_n}$  to be  $1, \dots, n$  respectively. In this way, we obtain the new set  $S_{new} = \{\hat{a}_1, \dots, \hat{a}_n\} \subset \mathbb{R}^d$  in  $O(n \log n)$  time (note that  $d$  is assumed to be constant). It is clear that  $S_{new}$  is regular. We verify that  $S_{new}$  satisfies the desired property. Assume  $a_i >_E a_j$ . Then in each dimension, the coordinate of  $a_i$  is greater than or equal to the coordinate of  $a_j$ . In addition, the sum of the  $d$  coordinates of  $a_i$  is greater than that of  $a_j$ . Therefore, in all dimensions, the coordinates of  $\hat{a}_i$  are greater than the coordinates of  $\hat{a}_j$ , i.e.,  $\hat{a}_i >_E \hat{a}_j$ . Assume  $a_i \not>_E a_j$ . Then there exists  $p \in \{1, \dots, d\}$  such that the  $p$ -th coordinate of  $a_i$  is smaller than the  $p$ -th coordinate of  $a_j$ . By definition, the  $p$ -th coordinate of  $\hat{a}_i$  is also smaller than the  $p$ -th coordinate of  $\hat{a}_j$ . Therefore,  $\hat{a}_i \not>_E \hat{a}_j$ .

## A.10 Proof of Lemma 18

Without loss of generality, we may assume  $B = E$ . Let  $[r_1 : \dots : r_d]$  be the homogeneous coordinates of  $l$ . Since  $l \notin PC_B$ , we may find  $r_p$  and  $r_q$  such that  $r_p > 0$  and  $r_q < 0$ . Now we define  $r'_1, \dots, r'_d \in \mathbb{R}$  by setting  $r'_p = -r_q$ ,  $r'_q = r_p$ , and  $r'_i = 0$  for any  $i \notin \{p, q\}$ . Consider the point  $x = [r'_1 : \dots : r'_d] \in \mathbb{P}^{d-1}$ . Note that  $r'_p$  and  $r'_q$  are nonzero so that  $x$  is well-defined. Since  $r'_1, \dots, r'_d$  are nonnegative, we have  $x \in PC_B$ . Furthermore, we know that  $\text{ang}(l, x) = \frac{\pi}{2}$ , because  $\sum_{i=1}^d r_i r'_i = 0$ .

## A.11 Proof of Lemma 19

Without loss of generality, we may assume  $B = E$ . Let  $[r_1 : \dots : r_d]$  be the homogeneous coordinates of  $x$  such that  $\sum_{i=1}^d r_i^2 = 1$ . Since  $x \in PC_B$ , the coordinates can be chosen such that  $r_1, \dots, r_d$  are nonnegative. Consider the point  $y = [r'_1 : \dots : r'_d] \in \mathbb{P}^{d-1}$  where  $r'_i = r_i + \frac{\varepsilon}{\sqrt{d}}$ . It is clear that  $y$  is well-defined and in  $PC_B$ . Set  $\theta = \text{ang}(x, y)$ . To see  $\theta < \varepsilon$ , we note that

$$\sin^2 \theta = 1 - \cos^2 \theta = 1 - \frac{(\sum_{i=1}^d r_i r'_i)^2}{\sum_{i=1}^d (r'_i)^2} = \frac{(d - \gamma^2)\varepsilon^2}{d + 2\varepsilon\sqrt{d}\gamma + d\varepsilon^2},$$

where  $\gamma = \sum_{i=1}^d r_i \geq 1$ . Therefore,  $\sin^2 \theta < \varepsilon^2/(1 + \varepsilon^2)$  and  $\sin \theta < \varepsilon/\sqrt{1 + \varepsilon^2}$ , which implies  $\theta < \varepsilon$ . It suffices to show that the  $\frac{\varepsilon}{3\sqrt{d}}$ -ball at  $y$  is contained in  $PC_B$ . Equivalently, we want that  $\text{ang}(z, y) > \frac{\varepsilon}{3\sqrt{d}}$  for any  $z \in \mathbb{P}^d \setminus PC_B$ . Let  $z = [s_1 : \dots : s_d]$  be a point in  $\mathbb{P}^d \setminus PC_B$  and assume  $\sum_{i=1}^d s_i^2 = 1$ . We have that

$$\cos^2(\text{ang}(z, y)) = \frac{(\sum_{i=1}^d r'_i s_i)^2}{\sum_{i=1}^d (r'_i)^2}.$$

Because  $z \in \mathbb{P}^d \setminus PC_B$ , there must exist some  $p, q$  such that  $s_p > 0$  and  $s_q < 0$ . Since  $r'_1, \dots, r'_d > 0$  and  $\sum_{i=1}^d s_i^2 = 1$ , we have that

$$\frac{(\sum_{i=1}^d r'_i s_i)^2}{\sum_{i=1}^d (r'_i)^2} \leq \frac{(\sum_{i=1}^d r'_i |s_i|)^2 - \eta^2}{\sum_{i=1}^d (r'_i)^2} \leq \frac{\sum_{i=1}^d (r'_i)^2 - \eta^2}{\sum_{i=1}^d (r'_i)^2},$$

where  $\eta = \min\{|r'_p|, |r'_q|\}$ . It follows that

$$\sin^2(\text{ang}(z, y)) \geq \frac{\eta^2}{\sum_{i=1}^d (r'_i)^2} \geq \frac{\varepsilon^2/d}{(1 + \varepsilon)^2} > \frac{\varepsilon^2}{9d}.$$

Therefore,  $\text{ang}(z, y) \geq \sin(\text{ang}(z, y)) > \frac{\varepsilon}{3\sqrt{d}}$ .

## A.12 Proof of Lemma 20

By taking  $\varepsilon > \frac{\pi}{2}$ , the statement in the theorem implies that for any  $\xi > 0$ , one can compute  $m = O(1/\xi^{d-1})$  points  $l_1, \dots, l_m \in \mathbb{P}^{d-1}$  in  $O(m)$  time such that  $\min_i \text{ang}(l_i, y) < \xi$  for any  $y \in \mathbb{P}^{d-1}$ . With this observation, we complete the proof by applying induction on the dimension. In  $\mathbb{P}^1$ , the statement is quite obvious. Without loss of generality, we may assume  $l = [0 : 1]$ . Set  $\gamma = \lfloor \varepsilon/\xi \rfloor$  and  $m = 2\gamma + 1$ . Then one can simply take the  $m$  points  $[\cos(i\xi) : \sin(i\xi)]$  for all  $i \in \{-\gamma, \dots, 0, \dots, \gamma\}$  as  $l_1, \dots, l_m$ . The two desired properties of  $l_1, \dots, l_m$  can be readily verified. Now suppose the theorem holds in  $\mathbb{P}^{k-1}$ , and we consider the case in  $\mathbb{P}^k$ . Similarly, we may assume  $l = [0 : \dots : 0 : 1] \in \mathbb{P}^k$ . As argued at the beginning, our induction hypothesis implies that we can compute  $m' = O(1/\xi^{k-1})$  points  $l'_1, \dots, l'_{m'} \in \mathbb{P}^{k-1}$  in  $O(m')$  time such that  $\min_i \text{ang}(l'_i, y) < \xi/2$  for any  $y \in \mathbb{P}^{k-1}$ . We then use these  $m'$  points to achieve our construction in  $\mathbb{P}^k$  as follows. For any real number  $\alpha \in [0, 1]$ , we define the inclusion map  $f_\alpha : \mathbb{P}^{k-1} \rightarrow \mathbb{P}^k$  as

$$f_\alpha : [r_1 : \dots : r_k] \mapsto \left[ r_1 : \dots : r_k : \sqrt{\frac{t}{1-\alpha^2}} \right],$$

where  $t = \sum_{i=1}^k r_i^2$  (note that  $f_\alpha$  is well-defined). Set  $\gamma = \lfloor 2\varepsilon/\xi \rfloor$  and  $m = (2\gamma + 1)m' = O(\varepsilon/\xi^k)$ . Also, set  $\alpha_i = \sin(i\xi/2)$  for  $i \in \{-\gamma, \dots, 0, \dots, \gamma\}$ . Then we take the  $m$  points  $f_{\alpha_i}(l'_j)$  for all  $i \in \{-\gamma, \dots, 0, \dots, \gamma\}$  and all  $j \in \{1, \dots, m'\}$  as  $l_1, \dots, l_m$ . It suffices to show that  $l_1, \dots, l_m$  satisfy the two desired conditions. Clearly, for any  $i \in \{-\gamma, \dots, 0, \dots, \gamma\}$  and  $j \in \{1, \dots, m'\}$ , we have that  $\text{ang}(l, f_{\alpha_i}(l'_j)) = \frac{\pi}{2} - i\xi/2 > \frac{\pi}{2} - \varepsilon$ . To verify the condition (2), let  $y = [r_1 : \dots : r_{k+1}]$  be a point in  $\mathbb{P}^k$  where  $\sum_{i=1}^{k+1} r_i^2 = 1$ . Suppose  $\text{ang}(l, y) > \frac{\pi}{2} - \varepsilon$ , so  $|r_{k+1}| < \sin \varepsilon$ . If  $r_{k+1} \geq 0$ , we define  $p$  as the largest integer in  $\{0, \dots, \gamma\}$  such that  $\sin(p\xi/2) \leq r_{k+1}$ , otherwise define  $p$  as the smallest integer in  $\{-\gamma, \dots, 0\}$  such that  $\sin(p\xi/2) \geq r_{k+1}$ . Set  $y' = [r_1 : \dots : r_k] \in \mathbb{P}^{k-1}$ . Then by assumption, there exists some  $q \in \{1, \dots, m'\}$  such that  $\text{ang}(l'_q, y') < \xi/2$ . We claim that  $\text{ang}(f_{\alpha_p}(l'_q), y) < \xi$ . Indeed, we consider the point  $f_{\alpha_p}(y') \in \mathbb{P}^k$ . We have  $\text{ang}(f_{\alpha_p}(l'_q), f_{\alpha_p}(y')) \leq \text{ang}(l'_q, y') < \xi/2$ . Also, we have  $\text{ang}(f_{\alpha_p}(y'), y) = |\arcsin(r_{k+1}) - p\xi/2| < \xi/2$ . Therefore,  $\text{ang}(f_{\alpha_p}(l'_q), y) < \xi$ , which implies that the points  $l_1, \dots, l_m$  satisfy the condition (2). The induction argument then completes the proof.

## A.13 Proof of Theorem 21

To prove the result, we first determine some constants. Since  $\mathcal{P}'$  is balanced, there is a constant  $c_1 < 1$  such that  $N - \max \Delta'_N \geq N^{c_1}$  for any sufficiently large  $N$ . Recall that our construction of the auxiliary point set  $\Psi$  satisfies  $|\Psi| = 2M = O(n^d)$  where  $n = |S|$ . So we can find a constant  $c_2$  such that  $|S \cup \Psi| \leq c_2 n^d$ . We construct the desired balanced color pattern  $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$  as follows. For an integer  $p > 0$ , in order to determine  $\Delta_p$ , set  $q = (c_2 p^d)^{2/c_1}$ . We consider two cases,  $|\Delta'_q| \geq c_2 p^d$  and  $|\Delta'_q| < c_2 p^d$ . In the case of  $|\Delta'_q| \geq c_2 p^d$ , we define  $\Delta_p = \{1, \dots, 1\}$ , i.e., a multi-set consisting of  $p$  1's. In the case of  $|\Delta'_q| < c_2 p^d$ , we define  $\Delta_p = \{\frac{p}{2}, \frac{p}{2}\}$  if  $p$  is even and  $\Delta_p = \{\frac{p-1}{2}, \frac{p+1}{2}\}$  if  $p$  is odd. This completes the construction of  $\mathcal{P}$ . We claim that the CSD problem in  $\mathbb{R}^d$  with respect to  $\mathcal{P}$  is polynomial-time reducible to the FBCSD problem in the same dimension with respect to  $\mathcal{P}'$ . Let  $\mathcal{S} = (S, \text{cl}, \pi)$  be a colored stochastic dataset in  $\mathbb{R}^d$  such that  $\langle \mathcal{S} \rangle$  is an instance of the CSD problem with respect to  $\mathcal{P}$ . Suppose  $|S| = n$  and set  $N = (c_2 n^d)^{2/c_1}$ . We want to construct another colored stochastic dataset  $\mathcal{S}' = (S', \text{cl}', \pi')$  in  $\mathbb{R}^d$  with  $|S'| = N$  such that  $\Lambda_{\mathcal{S}'}^* = \Lambda_{\mathcal{S}}$  and  $\langle \mathcal{S}' \rangle$  is an instance of the FBCSD problem with respect to  $\mathcal{P}'$ . As before, we first construct the auxiliary point set  $\Psi = \{b_1, \dots, b_M, b'_1, \dots, b'_M\}$ . By our assumption, we have  $|S \cup \Psi| = n + 2M \leq c_2 n^d < N$ . In order to have  $|S'| = N$ , we then arbitrarily choose a set  $D$  of  $N - (n + 2M)$  dummy points in  $\mathbb{R}^d$  (these points can be chosen arbitrarily as we will assign them existence probabilities 0 later) and set  $S' = S \cup \Psi \cup D$ . The existence probabilities of the points in  $S'$  are defined as

$$\pi'(a) = \begin{cases} \pi(a) & \text{if } a \in S, \\ 1 & \text{if } a \in \Psi, \\ 0 & \text{if } a \in D. \end{cases}$$

It suffices to define the coloring  $\text{cl}'$  of  $S'$ . Since we need  $\langle \mathcal{S}' \rangle$  to be an instance of the FBCSD problem with respect to  $\mathcal{P}'$ ,  $\text{cl}'$  must induce the partition  $\Delta'_N$ . Besides, it should be guaranteed that  $\text{cl}'(a) = \text{cl}(a)$

for all  $a \in S$  and  $\text{cl}'(b_i) \neq \text{cl}'(b'_i)$  for  $i \in \{1, \dots, M\}$  (as observed previously,  $\Lambda_{S'}^* = \Lambda_S$  as long as we have this). We consider two cases,  $|\Delta'_N| \geq c_2 n^d$  and  $|\Delta'_N| < c_2 n^d$ . In the case of  $|\Delta'_N| \geq c_2 n^d$ , we have  $\Delta_n = \{1, \dots, 1\}$  by definition and therefore all the points in  $S$  have distinct colors (under the coloring  $\text{cl}$ ). Note that  $|S \cup \Psi| \leq c_2 n^d \leq |\Delta'_N|$ . As such, one can easily find a coloring  $\text{cl}'$  inducing  $\Delta'_N$  which assigns distinct colors to the points in  $S \cup \Psi$  and satisfies  $\text{cl}'(a) = \text{cl}(a)$  for all  $a \in S$  (note that the coloring on  $D$  is “free”, so we can easily make  $\text{cl}'$  induces  $\Delta'_N$ ). This  $\text{cl}'$  completes our reduction. In the case of  $|\Delta'_N| < c_2 n^d$ , we have that  $\Delta_n = \{\frac{n}{2}, \frac{n}{2}\}$  if  $n$  is even and  $\Delta_n = \{\frac{n-1}{2}, \frac{n+1}{2}\}$  if  $n$  is odd. Without loss of generality, we may assume that  $\text{cl}(S) = \{1, 2\}$ . Suppose  $\Delta'_N = \{r_1, \dots, r_m\}$  where  $m < c_2 n^d$  and  $r_1 \geq \dots \geq r_m$ . We claim that  $r_1 \geq r_2 \geq c_2 n^d$ . Indeed, if  $r_2 < c_2 n^d$ , then  $\sum_{i=2}^m r_i < m c_2 n^d < (c_2 n^d)^2$  and hence  $N \leq (N - r_1)^{1/c_1} < (c_2 n^d)^{2/c_1}$ , contradicting the fact that  $N = (c_2 n^d)^{2/c_1}$ . With this observation, we try to construct  $\text{cl}'$  with  $\text{cl}'(S') = \{1, \dots, m\}$  such that  $\text{cl}'$  assigns color  $i$  to exactly  $r_i$  points in  $S'$ . We define  $\text{cl}'(a) = \text{cl}(a)$  for all  $a \in S$ ,  $\text{cl}'(b_i) = 1$  and  $\text{cl}'(b'_i) = 2$  for  $i \in \{1, \dots, M\}$ . Note that by doing this we do not “exhaust” the colors 1 and 2, because  $r_1 \geq r_2 \geq c_2 n^d \geq |S \cup \Psi|$ . So we can carefully determine  $\text{cl}'(a)$  for all  $a \in D$  such that exactly  $r_i$  points in  $S'$  have color  $i$ . By the defined  $\text{cl}'$ , we completes our reduction and the proof.

## A.14 Proof of Lemma 22

We first consider the “if” part. If  $L_R = \emptyset$ , then  $R$  is monochromatic and hence is good. If there exists  $l \in L_R$  such that  $\theta(l, l') > \frac{\pi}{2}$  for any  $l' \in L_R$  not equal to  $l$ , one can slightly rotate  $l$  clockwise to obtain  $l_0 \in \mathbb{P}^1$  such that  $\theta(l_0, l') > \frac{\pi}{2}$  for any  $l' \in L_R$ . Suppose the homogeneous coordinates of  $l_0$  is  $[\alpha : \beta]$  with  $\alpha^2 + \beta^2 = 1$ . Take the orthogonal basis  $B = (\mathbf{b}_1, \mathbf{b}_2)$  of  $\mathbb{R}^2$  with  $\mathbf{b}_1 = (\alpha, \beta)$  and  $\mathbf{b}_2 = (\beta, -\alpha)$ . Since  $\theta(l_0, l') > \frac{\pi}{2}$  for any  $l' \in L_R$ , we know that  $R$  contains no inter-color dominances with respect to  $B$ . To see the “only if” part, let  $R$  be a good realization of  $\mathcal{S}$ . Suppose  $R$  contains no inter-color dominances with respect to some orthogonal basis  $B = (\mathbf{b}_1, \mathbf{b}_2)$  of  $\mathbb{R}^2$  (assume  $\mathbf{b}_2$  is in the clockwise direction of  $\mathbf{b}_1$  with angle  $\frac{\pi}{2}$ ). Let  $b \in \mathbb{P}^1$  be the point corresponding to  $\mathbf{b}_1$  (i.e.,  $b$  is the image of  $\mathbf{b}_1$  under the obvious quotient map  $S^1 \rightarrow \mathbb{P}^1$ ). If  $L_R = \emptyset$ , we are done. So assume  $L_R \neq \emptyset$ . Define  $l \in L_R$  as the point which minimizes  $\theta(l, b)$ . We claim that  $\theta(l, l') > \frac{\pi}{2}$  for any  $l' \in L_R$  not equal to  $l$ . Let  $l' \in L_R$  be a point not equal to  $l$ . If  $\theta(l, l') \leq \frac{\pi}{2}$ , then either  $\theta(l', b) < \theta(l, b)$  or  $l' \in PC_B$  (recall that  $PC_B$  is the projective cone of  $B$  defined in Section 3.1). The former contradicts the definition of  $l$  while the latter contradicts the fact that  $R$  contains no inter-color dominances with respect to  $B$ .

## A.15 Proof of Lemma 23

First, we observe (Observation 1 hereafter) that a realization  $R$  of  $\mathcal{S}$  is good with  $\text{wit}(R) = (a_{i*}, a_{j*})$  iff (1)  $a_{i*}, a_{j*} \in R$  and (2) for any  $a_i, a_j \in R$  with  $\text{cl}(a_i) \neq \text{cl}(a_j)$  and  $a_i >_B a_j$ , we have  $\max(i, j) \leq j^*$  and  $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$ . To see the “if” part, assume  $R$  satisfies the conditions (1) and (2). Since  $a_{i*}, a_{j*} \in R$ , we know that  $\overline{a_{i*} - a_{j*}} \in L_R$ . Set  $l = \overline{a_{i*} - a_{j*}}$ . The condition (2) guarantees that  $\theta(l, l') > \frac{\pi}{2}$  for any  $l' \in L_R$  not equal to  $l$ . Thus, by Lemma 22,  $R$  is good (but not monochromatic since  $a_{i*}, a_{j*} \in R$ ). Furthermore, it is easy to see that the conditions (1) and (2) also guarantee  $\text{wit}(R) = (a_{i*}, a_{j*})$ . To see the “only if” part, assume  $R$  is good with  $\text{wit}(R) = (a_{i*}, a_{j*})$ . By the definition of witness pair, we immediately have  $a_{i*}, a_{j*} \in R$ . Again, set  $l = \overline{a_{i*} - a_{j*}}$ . Let  $a_i, a_j \in R$  be two points such that  $\text{cl}(a_i) \neq \text{cl}(a_j)$  and  $a_i >_B a_j$ . By the definition of  $B$ , we have  $\theta(l, l') \leq \frac{\pi}{2}$  for  $l' = \overline{a_i - a_j}$ . According to Lemma 22, it implies that  $l = l'$ , i.e.,  $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$ . Besides, we must have  $\max(i, j) \leq j^*$ , otherwise  $(a_{i*}, a_{j*})$  is not the witness pair of  $R$ .

Second, we observe (Observation 2 hereafter) that our construction of  $\mathcal{S}'$  satisfy the following property. Let  $i, j \in \{1, \dots, n\}$  be any indices such that  $\text{cl}(a_i) \neq \text{cl}(a_j)$ , or equivalently,  $\text{cl}'(a'_i) \neq \text{cl}'(a'_j)$ . Then we have  $a'_i >_E a'_j$  iff (1)  $a_i >_B a_j$  and (2)  $\langle \mathbf{b}_2, a_i \rangle > \langle \mathbf{b}_2, a_j \rangle$  or  $\max(i, j) > j^*$ . To see the “if” part, assume  $a_i >_B a_j$ . In this case, we have  $y(a'_i) = \langle \mathbf{b}_1, a_i \rangle \geq \langle \mathbf{b}_1, a_j \rangle = y(a'_j)$ . If  $\langle \mathbf{b}_2, a_i \rangle > \langle \mathbf{b}_2, a_j \rangle$ , then  $x(a'_i) \geq \langle \mathbf{b}_2 - \delta, a_i \rangle > \langle \mathbf{b}_2, a_j \rangle \geq x(a'_j)$  so that  $a'_i >_E a'_j$ . If  $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$  and  $\max(i, j) > j^*$ , we also have  $x(a'_i) = \langle \mathbf{b}_2, a_i \rangle > \langle \mathbf{b}_2, a_j \rangle = x(a'_j)$  (recall the general position assumption) so that  $a'_i >_E a'_j$ . To see the “only if” part, first assume  $a_i \not>_B a_j$ . In this case, we have either  $y(a'_i) < y(a'_j)$  or  $x(a'_i) < x(a'_j)$ , which



implies  $a'_i \not\prec_E a'_j$ . Now assume  $a_i >_B a_j$ ,  $\langle \mathbf{b}_2, a_i \rangle \leq \langle \mathbf{b}_2, a_j \rangle$ , and  $\max(i, j) \leq j^*$ . Because  $a_i >_B a_j$ , it must be the case that  $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$  and  $\langle \mathbf{b}_1, a_i \rangle > \langle \mathbf{b}_1, a_j \rangle$ . By our construction, we have  $x(a'_i) = \langle \mathbf{b}_2, a_i \rangle - \delta$ . But  $x(a'_j) = \langle \mathbf{b}_2, a_j \rangle$  (recall the general position assumption). Thus,  $a'_i \not\prec_E a'_j$ .

With the above two observations, we prove the equation  $Pr_{i^*, j^*} = \pi(a_{i^*}) \cdot \pi(a_{j^*}) \cdot \Gamma_{S'}$ . Define a natural one-to-one correspondence  $\mu : S \rightarrow S'$  as  $\mu(a_i) = a'_i$ . First, it is clear that for any subset  $A \subseteq S$  including  $a_{i^*}, a_{j^*}$ , the probability that  $A$  occurs as a realization of  $\mathcal{S}$  is equal to the product  $\pi(a_{i^*}) \cdot \pi(a_{j^*}) \cdot \Pr[\mu(A)]$ , where  $\Pr[\mu(A)]$  the probability that  $\mu(A)$  occurs as a realization of  $S'$ . Let  $R$  be a realization of  $\mathcal{S}$ . We claim that  $R$  is good with  $\text{wit}(R) = (a_{i^*}, a_{j^*})$  iff  $a'_{i^*}, a'_{j^*} \in \mu(R)$  and  $\mu(R)$  contains no inter-color dominances (with respect to  $E$ ). To see the “if” part, assume  $a'_{i^*}, a'_{j^*} \in \mu(R)$  and  $\mu(R)$  contains no inter-color dominances with respect to  $E$ . Then  $a_{i^*}, a_{j^*} \in R$ . Let  $a_i, a_j \in R$  be two points such that  $\text{cl}(a_i) \neq \text{cl}(a_j)$  and  $a_i >_B a_j$ . Since  $\mu(R)$  contains no inter-color dominances with respect to  $E$ , Observation 2 above implies that  $\max(i, j) \leq j^*$  and  $\langle \mathbf{b}_2, a_i \rangle \leq \langle \mathbf{b}_2, a_j \rangle$  (the latter further implies  $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$  since  $a_i >_B a_j$ ). Thus, by Observation 1 above,  $R$  is good with  $\text{wit}(R) = (a_{i^*}, a_{j^*})$ . To see the “only if” part, assume  $R$  is good with  $\text{wit}(R) = (a_{i^*}, a_{j^*})$ . Then Observation 1 implies  $a_{i^*}, a_{j^*} \in R$  and hence  $a'_{i^*}, a'_{j^*} \in \mu(R)$ . Let  $a_i, a_j \in R$  be two points such that  $\text{cl}(a_i) \neq \text{cl}(a_j)$ . If  $a_i \not\prec_B a_j$ , then by Observation 2 we have  $a'_i \not\prec_E a'_j$ . If  $a_i >_B a_j$ , then Observation 1 implies that  $\max(i, j) \leq j^*$  and  $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$ . By using Observation 2, we also have  $a'_i \not\prec_E a'_j$ . Therefore,  $\mu(R)$  contains no inter-color dominances with respect to  $E$ . This argument shows that  $\mu$  induces a one-to-one correspondence between the good realizations of  $\mathcal{S}$  and the realizations of  $S'$  which include  $a'_{i^*}, a'_{j^*}$  and contain no inter-color dominances (with respect to  $E$ ). Note that  $\pi'(a'_{i^*}) = \pi'(a'_{j^*}) = 1$ , hence a realization of  $S'$  for sure includes  $a'_{i^*}, a'_{j^*}$ . As a result, we have  $Pr_{i^*, j^*} = \pi(a_{i^*}) \cdot \pi(a_{j^*}) \cdot \Gamma_{S'}$ .

## B Solving the CSD problem for $d = 2$ in $O(n^2 \log^2 n)$ time

In this section, we give the details of our improved algorithm using 2D range trees. Formally, the 2D range tree,  $\mathcal{T}$ , used in this paper is built on a fixed collection of planar points and maintains the weights of these points. It supports the following three operations:

- $\text{QUERY}(\mathcal{T}, r)$  returns the sum of weights of all the points in the query range  $r$ .
- $\text{UPDATE}(\mathcal{T}, p, w)$  updates the weight of point  $p$  to  $w$ .
- $\text{MULTIPLY}(\mathcal{T}, r, \delta)$  multiplies by a factor of  $\delta$  the weight of every point in the range  $r$ . Note that this operation is revertible and the inverse of  $\text{MULTIPLY}(\mathcal{T}, r, \delta)$  is  $\text{MULTIPLY}(\mathcal{T}, r, 1/\delta)$ .

With a careful implementation, see Appendix B.1, all three operations can run in  $O(\log^2 n)$  time.

Two more notations are defined. For a legal pair  $(i, j)$ , we use  $(i, j)_{\searrow}$  (resp.  $(i, j)_{\swarrow}$ ) to represent the point  $(x(a_i), y(a_j))$  (resp.  $(x(a_j), y(a_i))$ ); see Figure 11. Also, let  $\text{QUAD}(p)$  denote the northwest open quadrant of point  $p$ , i.e.,  $(-\infty, x(p)) \times (y(p), \infty)$ . We now give the complete solution shown in Algorithm 1 followed by the correctness analysis.

---

**Algorithm 1** Computing  $\Gamma_{\mathcal{S}}$  in  $O(n^2 \log^2 n)$  time.

---

```

1: procedure COMPUTE- $\Gamma_{\mathcal{S}}(\mathcal{S})$   $\triangleright$  Recall  $\mathcal{S} = (S, \text{cl}, \pi)$ .
2:   Sort all points in  $S$  such that  $x(a_1) < \dots < x(a_n)$ .
3:   Let  $\mathcal{T}$  be the 2D range tree built on  $\{(i, j)_{\searrow} : (i, j) \text{ is legal}\}$  with initial weights 0.
4:   Let  $\mathcal{T}_k$  be the 2D range tree built on  $\{(i, j)_{\searrow} : (i, j) \text{ is legal and } \text{cl}(a_i) = \text{cl}(a_j) = k\}$  with initial
   weights 0, for every color  $k$ .
5:    $prod = \prod_{i=1}^n (1 - \pi(a_i))$ 
6:    $\Gamma_{\mathcal{S}} \leftarrow prod$ 
7:    $\text{UPDATE}(\mathcal{T}, a_0, 1)$   $\triangleright$  This implies  $F(0, 0) = 1$ . Also, no need to update  $\mathcal{T}_{\text{cl}(a_0)}$ .
8:   for  $i \leftarrow 1$  to  $n$  do
9:      $prod \leftarrow prod \cdot (1 - \pi(a_i))^{-1}$ 
10:     $k \leftarrow \text{cl}(a_i)$ 
11:     $\text{MULTIPLY}(\mathcal{T}, \text{QUAD}(a_j), (1 - \pi(a_j))^{-1})$  and  $\text{MULTIPLY}(\mathcal{T}_k, \text{QUAD}(a_j), (1 - \pi(a_j))^{-1})$  for all  $j \in$ 
     $\{1, \dots, i\}$  such that  $\text{cl}(a_j) = k$ .
12:    Let  $(\ell_1, \dots, \ell_i)$  be a permutation of  $(1, \dots, i)$  such that  $y(a_{\ell_1}) < \dots < y(a_{\ell_i})$ .
13:    for  $j \leftarrow \ell_1$  to  $\ell_i$  do
14:      if  $(i, j)$  is a legal pair then  $\triangleright$  This implies that  $\text{cl}(a_j) = k$ .
15:         $F(i, j) \leftarrow \text{QUERY}(\mathcal{T}, \text{QUAD}((i, j)_{\swarrow})) - \text{QUERY}(\mathcal{T}_k, \text{QUAD}((i, j)_{\swarrow}))$ 
16:         $F(i, j) \leftarrow F(i, j) \cdot \pi_{i,j}^*$ 
17:         $\Gamma_{\mathcal{S}} \leftarrow \Gamma_{\mathcal{S}} + F(i, j) \cdot prod$ 
18:         $\text{MULTIPLY}(\mathcal{T}, \text{QUAD}(a_j), 1 - \pi(a_j))$ 
19:         $\text{MULTIPLY}(\mathcal{T}_k, \text{QUAD}(a_j), 1 - \pi(a_j))$ 
20:      end if
21:    end for
22:    Revert all MULTIPLY operations executed in Line 11, 18, 19.
23:     $\text{UPDATE}(\mathcal{T}, (i, j)_{\searrow}, F(i, j))$  and  $\text{UPDATE}(\mathcal{T}_k, (i, j)_{\searrow}, F(i, j))$  for every  $j \in \{1, \dots, i\}$  such that pair
     $(i, j)$  is legal.
24:     $\text{MULTIPLY}(\mathcal{T}, (-\infty, x(a_i)) \times \mathbb{R}, 1 - \pi(a_i))$ 
25:     $\text{MULTIPLY}(\mathcal{T}_k, (-\infty, x(a_i)) \times \mathbb{R}, 1 - \pi(a_i))$ 
26:  end for
27:  return  $\Gamma_{\mathcal{S}}$ 
28: end procedure

```

---

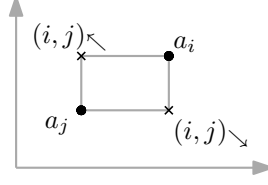


Figure 11: Illustrating  $(i, j)_{\nwarrow}$  and  $(i, j)_{\nearrow}$  for a legal pair  $(i, j)$ .

**Correctness analysis.** We compute  $F(i, j)$  for each legal pair  $(i, j)$  by first enumerating  $i$  from 1 to  $n$  and then  $j$  in an order such points are visited from bottom to top; see the nested loop at Line 8 and 13. For now, assume the fact, which we prove later, that the inner  $j$ -loop correctly computes  $F(i, j)$  for all legal pairs  $(i, j)$  when  $i$  is fixed. We then have the following lemma.

**Lemma 25** *At the beginning of the  $i$ -th iteration (Line 10), the weight of  $(i', j')_{\nwarrow}$  in  $\mathcal{T}$ , such that  $i' < i$ , is equal to  $F(i', j') \cdot \prod_{p \in S \cap \parallel} (1 - \pi(p))$ , where  $\parallel$  denotes the open strip  $(x(a_{i'}), x(a_i)) \times \mathbb{R}$ . (See Figure 12a.)*

*Proof.* This statement is trivially true for  $i = 1$  as all the weights in  $\mathcal{T}$  are equal to zero except that  $F(0, 0) = 1$ . Assume the statement is true for the  $i$ -th iteration, we show it also holds for the  $(i + 1)$ -th iteration. First, we can safely consider  $\mathcal{T}$  unchanged throughout Line 10-22 because although Line 11 and 18 modifies  $\mathcal{T}$ , these side-effects are reverted immediately in Line 22. After the inner  $j$ -loop is done, by our early assumption, we obtain the value of  $F(i, j)$  for every legal pair  $(i, j)$  when  $i$  is fixed. These values are not currently stored in  $\mathcal{T}$  but are needed for the next iteration. Thus, we update the weight of each  $(i, j)_{\nwarrow} \in \mathcal{T}$  to  $F(i, j)$ , as stated in Line 23. We also need to multiply the factor  $(1 - \pi(a_i))$  to the weight of each  $(i', j')_{\nwarrow} \in \mathcal{T}$  that is to the left of  $a_i$  because  $a_i$  will be included in the strip as we proceed from  $i$  to  $i + 1$ . This is handled by Line 24. As such, the statement is maintained for the  $(i + 1)$ -th iteration, which completes the proof.  $\square$

With Lemma 25 in hand, we now give the proof of our aforementioned statement, as restated in Lemma 26.

**Lemma 26** *Line 15-16 correctly computes  $F(i, j)$ .*

*Proof.* Recall that  $F(i, j) = \pi_{i,j}^* \cdot \sum_{(i', j') \in J_{i,j}} F(i', j') \cdot \Pi_{i,j,i',j'}$ . By Lemma 25, at the beginning of the  $i$ -th round, the weight of each  $(i', j')_{\nwarrow} \in \mathcal{T}$ , where  $i' < i$ , is equal to  $F(i', j') \cdot \prod_{p \in S \cap \parallel} (1 - \pi(p))$ . This product is off from the ideal one,  $F(i', j') \cdot \Pi_{i,j,i',j'}$ , by a factor of  $\prod_{p \in S^{(i)} \cap \square} (1 - \pi(p))$ , where  $S^{(i)} = \{p \in S : \text{cl}(p) = \text{cl}(a_i)\}$  and  $\square$  denotes the box  $(x(a_{i'}), x(a_i)) \times [y(a_j), y(a_{j'})]$ ; see Figure 12b. To cancel this factor, we observe that

$$\prod_{p \in S^{(i)} \cap \square} (1 - \pi(p)) = \prod_{p \in S^{(i)} \cap \sqcap_1} (1 - \pi(p)) \Bigg/ \prod_{p \in S^{(i)} \cap \sqcap_2} (1 - \pi(p)),$$

where  $\sqcap_1$  and  $\sqcap_2$  respectively denote the three-sided rectangle  $(x(a_{i'}), x(a_i)) \times (-\infty, y(a_{j'}))$  and  $(x(a_{i'}), x(a_i)) \times (-\infty, y(a_j))$ ; see Figure 12c and 12d. The former product ( $\sqcap_1$ ) is canceled in Line 11, and the latter ( $\sqcap_2$ ) is gradually accumulated back via  $(j - 1)$  calls of Line 18 as  $a_{\ell_1}, \dots, a_{\ell_{j-1}}$  are all below  $a_{\ell_j}$ . Thus, the weight of each  $(i', j')_{\nwarrow} \in \mathcal{T}$  is equal to  $F(i', j') \times \Pi_{i,j,i',j'}$  right before  $F(i, j)$  gets evaluated. Finally, the range query in Line 15 sums up the weight of every  $(i', j')_{\nwarrow} \in \mathcal{T}$  such that  $(i', j') \in J_{i,j}$ . (Note that the subtraction in Line 15 is needed because  $\text{QUERY}(\mathcal{T}, \text{QUAD}((i, j)_{\nwarrow}))$  also counts the probabilities of those legal pairs that have the same color as  $\text{cl}(a_i)$ .) Therefore, the value of  $F(i, j)$  is correctly computed after Line 16.  $\square$

Though Lemma 25 and 26 are cross-referencing, one can easily figure out that this is not a circular reasoning and is indeed a valid proof. Also, both lemmas can directly apply to  $\mathcal{T}_k$ 's as we always query/update  $\mathcal{T}$  and  $\mathcal{T}_k$ 's in the same way. Finally, all  $F(i, j)$ 's are computed and added up into  $\Gamma_S$ , which completes the correctness proof of the entire algorithm.

The overall runtime of Algorithm 1 is  $O(n^2 \log^2 n)$  since there are  $O(n^2)$  range queries/updates, each of which takes  $O(\log^2 n)$  time. The space occupied by  $\mathcal{T}$ , denoted by  $|\mathcal{T}|$ , is  $O(n^2 \log n^2) = O(n^2 \log n)$  as there

are  $O(n^2)$  legal pairs. Similarly, let  $n_k$  be the number of points in color  $k$ , and then  $\mathcal{T}_k$  costs  $O(n_k^2 \log n_k)$  space. Assume there are  $K$  colors in total. We have  $n_1 + \dots + n_K = n$  and thus  $|\mathcal{T}_1| + \dots + |\mathcal{T}_K| = O(n^2 \log n)$ . The overall space complexity is  $O(|\mathcal{T}| + |\mathcal{T}_1| + \dots + |\mathcal{T}_K|) = O(n^2 \log n)$ .

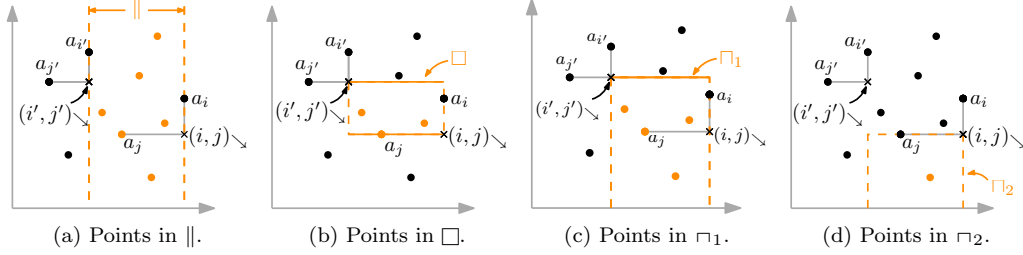


Figure 12: Illustrating Lemma 26. Orange color is only used to highlight each range and does *not* represent the color of each point. Dashed (resp. solid) boundaries are exclusive (resp. inclusive).

## B.1 Implementation details of our 2D range tree

In this section, we discuss how to implement an augmented 2D range tree,  $\mathcal{T}$ , to dynamically support QUERY, UPDATE, and MULTIPLY in  $O(\log^2 m)$  time, where  $m$  is the input size.

We first describe how to implement a dynamic 1D range tree,  $\mathcal{T}_{1D}$ , built on the  $y$ -coordinates of a set of planar points,  $P$ , to support the three operations, where the range used in QUERY and MULTIPLY is a 1D interval. The leaves, sorted by increasing  $y$ -coordinates, of  $\mathcal{T}_{1D}$  are points in  $P$  with initial weight equal to 0. In addition, in each internal node,  $u$ , we store two fields,  $sum(u)$  and  $mul(u)$ , where the former is the sum of weights in the subtree rooted at  $u$  and the latter is the multiplication-factor that needs to be applied to all the nodes in the subtree. For simplicity we use the notion  $sum(u)$  to denote the weight of  $u$  if it is a leaf. Also, set  $sum(u) = 0$  and  $mul(u) = 1$  initially.

Given a query/update range, we first identify  $O(\log m)$  canonical nodes,  $\mathcal{C}$ , of  $\mathcal{T}_{1D}$  via a recursive down-phase traversal. We then aggregate or modify the data in each canonical node. Finally, we refine the fields of those nodes along the path from every canonical node up to the root, as the recursion gradually terminates.

In the down-phase, when a non-leaf node  $u$  is visited, we call the following PUSH method to revise  $sum(u)$  based on  $mul(u)$  and then push the factor further to its two children. In the up-phase, we apply the COMBINE method to each node to readjust the sum. Between the down and up phase, we perform one of the following three operations.

- Add up  $sum(u)$  for every  $u \in \mathcal{C}$  for QUERY( $\mathcal{T}_{1D}, p$ ).
- Update  $sum(u)$  to  $w$  for the *only* element  $u \in \mathcal{C}$  for UPDATE( $\mathcal{T}_{1D}, p, w$ ).
- Multiply  $mul(u)$  by a factor of  $\delta$  for every  $u \in \mathcal{C}$  for MULTIPLY( $\mathcal{T}_{1D}, p, \delta$ ).

Finally, we build our 2D range tree,  $\mathcal{T}$ , on the  $x$ -coordinates of the given input. For each node  $u \in \mathcal{T}$ , we build an aforementioned 1D range tree w.r.t. the set of points in  $u$ . We also store at  $u$  a similar tag indicating the multiplication factor that needs to be applied to the 1D range tree stored at  $u$  as well as all  $u$ 's descendants. Given a 2D range query, we do a down-phase traversal identifying  $O(\log m)$  canonical nodes of  $\mathcal{T}$ . For each visited node  $u$  during the traversal, we should apply the multiplication tag to the 1D tree stored at  $u$  and push it further to  $u$ 's two children. This takes  $O(\log m)$  time. Then, for every canonical node  $u$ , we spend another  $O(\log m)$  time querying the 1D range tree stored at  $u$ , as stated above. Therefore, all three operations can be done in  $O(\log^2 m)$  time, and  $\mathcal{T}$  occupies  $O(m \log m)$  space.

---

**Algorithm 2** Implementation details of PUSH and COMBINE.

---

```
1: procedure PUSH( $u$ ) ▷ Only called in the down-phase.
2:    $sum(u) \leftarrow sum(u) \cdot mul(u)$ 
3:   if  $u$  is not a leaf then
4:      $mul(lchild(u)) \leftarrow mul(lchild(u)) \cdot mul(u)$ 
5:      $mul(rchild(u)) \leftarrow mul(rchild(u)) \cdot mul(u)$ 
6:   end if
7:    $mul \leftarrow 1$ 
8: end procedure
9: procedure COMBINE( $u$ ) ▷ Only called in the up-phase and we must have  $mul(u) = 1$ .
10:   $sum(u) \leftarrow sum(lchild(u)) + sum(rchild(u))$ 
11: end procedure
```

---

## B.2 Handling range-multiplication/division with a factor of zero

One may notice that the implementation above contains a flaw for  $MULTIPLY(\mathcal{T}, r, \delta)$  when  $\delta = 0$  because the inverse of this operation does not exist as  $1/0$  is undefined. We can overcome this issue by adding in each node a *zero-counter* and counting the number of zero factors separately. That is, if  $MULTIPLY$  multiplies a factor of zero, we increment the zero-counter of each canonical node instead of modifying  $sum$  and  $mul$  fields; if  $MULTIPLY$  divides a factor of zero, we decrement the corresponding zero-counters. Also, when a  $QUERY$  is triggered, we simply return zero for those canonical nodes whose zero-counter is positive. This solves the problem without increasing the runtime of all three operations.

## C A generalization of Lemma 6

In this section, we extend Lemma 6 to a general result revealing the hardness of stochastic geometric problems (under existential uncertainty). Many stochastic geometric problems focus on computing the probability that a realization of the given stochastic dataset has some specific property, e.g., [3, 5, 8, 18] and this paper. This kind of problems can be abstracted and generalized as follows. Let  $\mathcal{C}$  be a category of geometric objects (e.g., points, lines, etc.), and  $\mathbf{P}$  be a property defined on finite sets of objects in  $\mathcal{C}$ .

**Definition 27** *We define the  $\mathbf{P}$ -probability-computing problem as follows. The input is a stochastic dataset  $\mathcal{S} = (S, \pi)$  where  $S$  is a set of objects in  $\mathcal{C}$  and  $\pi : S \rightarrow (0, 1]$  is the function defining existence probabilities for the objects. The goal is to compute the probability that a realization of  $\mathcal{S}$  has the property  $\mathbf{P}$ .*

**Example 1.** Let  $\mathcal{C}$  be the category of points in  $\mathbb{R}^d$ , and  $\mathbf{P}$  be the property that the convex-hull of the set of points (in  $\mathbb{R}^d$ ) contains a fixed point  $q \in \mathbb{R}^d$ . In this case, the  $\mathbf{P}$ -probability-computing problem is the convex-hull membership probability problem [3].

**Example 2.** Let  $\mathcal{C}$  be the category of bichromatic points in  $\mathbb{R}^d$ , and  $\mathbf{P}$  be the property that the set of bichromatic points is linearly separable. In this case, the  $\mathbf{P}$ -probability-computing problem is the stochastic linear separability problem [5, 18].

**Example 3.** Let  $\mathcal{C}$  be the category of points in  $\mathbb{R}^d$ , and  $\mathbf{P}$  be the property that the closest-pair distance of the set of points is at most a fixed threshold  $\ell$ . In this case, the  $\mathbf{P}$ -probability-computing problem is the stochastic closest-pair problem [8].

By generalizing Definition 5, we may also consider an abstract notion of the cardinality-sensitive-counting problem.

**Definition 28** *Let  $c$  be a fixed constant. We define the  $\mathbf{P}$ -cardinality-sensitive-counting problem as follows. The input consists of a set  $S$  of objects in  $\mathcal{C}$  and a  $c$ -tuple  $(S_1, \dots, S_c)$  of disjoint subsets of  $S$ . The goal is to compute, for every  $c$ -tuple  $(n_1, \dots, n_c)$  of integers where  $0 \leq n_i \leq |S_i|$ , the number of the subsets  $S' \subseteq S$  which have the property  $\mathbf{P}$  and satisfy  $|S' \cap S_i| = n_i$  for all  $i \in \{1, \dots, c\}$ .*

**Example 4.** Consider the following problem: given a set  $S$  of bichromatic (red/blue) points in  $\mathbb{R}^d$ , compute the number of the linearly separable subsets  $S' \subseteq S$  which contain an equal number of red and blue points. This is clearly a restricted version of the  $\mathbf{P}$ -cardinality-sensitive-counting problem, where  $\mathcal{C}$  is the category of bichromatic points in  $\mathbb{R}^d$  and  $\mathbf{P}$  is the property that the set of bichromatic points is linearly separable.

The following theorem, which generalizes Lemma 6, implies that the  $\mathbf{P}$ -probability-computing problem is at least as “hard” as the  $\mathbf{P}$ -cardinality-sensitive-counting problem. The proof is (almost) the same as that of Lemma 6 (see Appendix A.2).

**Theorem 29** *The  $\mathbf{P}$ -cardinality-sensitive-counting problem is polynomial-time reducible to the  $\mathbf{P}$ -probability-computing problem for any  $\mathbf{P}$ .*

## D Implication in order dimension theory

In Section 2.2.3, we achieved the result that  $\dim(G) \leq 7$  for any 3-regular planar bipartite graph  $G$ . In this section, we establish an implication of this result in order dimension theory [15]. Let  $X$  be a finite set and  $<_P$  be a partial order on  $X$ . A set  $\{<_1, \dots, <_t\}$  of linear orders (or total orders) on  $X$  is said to be a *realizer* of  $<_P$  if

$$<_P = \bigcap_{i=1}^t <_i,$$

that is, for any  $x, y \in X$ ,  $x <_P y$  iff  $x <_i y$  for all  $i \in \{1, \dots, t\}$ . The *order dimension*  $\dim(<_P)$  of  $<_P$  is defined as the least cardinality of a realizer of  $<_P$  (see for example [15]).

The partial order  $<_P$  can be represented by a transitive directed graph  $G_{<_P} = (X, E)$  where  $E = \{\langle x, y \rangle : x <_P y\}$ . The *comparability graph*  $H_{<_P}$  of  $<_P$  is defined as the underlying undirected graph of  $G_{<_P}$ , i.e.,  $H_{<_P} = (X, E')$  where  $E' = \{(x, y) : x <_P y\}$ . Our result implies the following.

**Corollary 30** *Let  $(X, <_P)$  be a partial ordered set. If the comparability graph  $H_{<_P}$  is 3-regular planar bipartite, then  $\dim(<_P) \leq 7$ .*

*Proof.* Suppose  $H_{<_P} = (X_1 \cup X_2, E)$ , which is 3-regular planar bipartite. We must construct a realizer of  $<_P$  of size at most 7. Without loss of generality, we may assume  $H_{<_P}$  is connected (otherwise we could work on each connected components separately). Using our result in Section 2.2.3, we have  $\dim(H_{<_P}) \leq 7$ , so there exists a DPE  $f : X_1 \cup X_2 \rightarrow \mathbb{R}^7$  of  $H_{<_P}$ . Let  $Y$  be the image of  $f$ . By Lemma 15, we may further assume that  $Y$  is regular in  $\mathbb{R}^7$  (see Section 3.1 for definition). Now define a directed graph  $G_f = (X, E_f)$  as  $E_f = \{\langle x, y \rangle : f(y) > f(x)\}$ . It is clear that  $G_f$  is transitive and  $H_{<_P}$  is the underlying undirected graph of  $G_f$ . Since  $H_{<_P}$  is connected, the edges in  $G_f$  must be all directed from  $X_1$  to  $X_2$  or all directed from  $X_2$  to  $X_1$  (otherwise  $G_f$  is not transitive). On the other hand,  $H_{<_P}$  is also the underlying undirected graph of  $G_{<_P}$  (defined above) and  $G_{<_P}$  is also transitive. Therefore, either  $G_f = G_{<_P}$  or  $G_f$  and  $G_{<_P}$  are “reverses” of each other ( $G_f$  is the same as  $G_{<_P}$  except the orientations of the edges are reversed). If  $G_f = G_{<_P}$ , we define a set  $\{<_1, \dots, <_7\}$  of linear orders on  $X$  as  $x <_i y$  iff the  $i$ -th coordinate of  $f(x)$  is smaller than  $i$ -th coordinate of  $f(y)$ . If  $G_f$  and  $G_{<_P}$  are reverses of each other, we define  $\{<_1, \dots, <_7\}$  as  $x <_i y$  iff the  $i$ -th coordinate of  $f(x)$  is greater than  $i$ -th coordinate of  $f(y)$ . Since  $Y$  is regular,  $<_1, \dots, <_7$  are truly linear orders on  $X$ . It suffices to verify that  $<_P = \bigcap_{i=1}^7 <_i$ . We only verify for the case of  $G_f = G_{<_P}$ , the other case is similar. Suppose  $x <_P y$ . Then  $\langle x, y \rangle$  is an edge of  $G_{<_P}$  and also an edge of  $G_f$ . By the definition of  $G_f$ , we have  $f(y) > f(x)$ , which implies  $x <_i y$  for all  $i \in \{1, \dots, 7\}$ . Suppose  $x <_i y$  for all  $i \in \{1, \dots, 7\}$ . Then  $f(y) > f(x)$ . Hence,  $\langle x, y \rangle$  is an edge of  $G_f$  and also an edge of  $G_{<_P}$ . It follows that  $x <_P y$ .  $\square$